

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

Shivam Kumaran<sup>1\*</sup>, Samir Mandal<sup>1</sup>, Sudip Bhattacharyya<sup>2</sup>, Deepak Mishra<sup>3</sup>

<sup>1</sup>Department of Earth and Space Sciences, Indian Institute of Space Science and Technology, Thiruvananthapuram, 695547, India

<sup>2</sup>Department of Astronomy and Astrophysics, Tata Institute of Fundamental Research, Mumbai, 400005, India

<sup>3</sup>Department of Avionics, Indian Institute of Space Science and Technology, Thiruvananthapuram, 695547, India

Accepted XXX. Received YYY; in original form ZZZ

## ABSTRACT

Many sky surveys have found a large number of unidentified sources, which 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), low-mass X-ray binaries (LMXBs), high-mass X-ray binaries (HMXBs), ultra luminous X-ray sources (ULXs), cataclysmic variables (CVs), and pulsars. The catalogue consists of  $\approx 317000$  sources, out of which we select 277069 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 *Spitzer*'s MIPS, 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% accuracy, 81% precision, 79% recall, 80% f1 score and 0.91 Mathew's Correlation coefficient score. With the trained classifier, we identify 54770 (14066) sources with more than  $3\sigma$  ( $4\sigma$ ) confidence, out of which there are 32600 (8574) AGN, 16148 (5166) stars, 5184 (208) YSOs, 439 (46) HMXBs, 197 (71) LMXBs, 50 (0) ULXs, 89 (1) CVs, and 63 (0) pulsars. Our classifier can be useful to reliably identify sources of other catalogues.

**Key words:** keyword1 – keyword2 – keyword3

## 1 INTRODUCTION

With the advent of advanced X-ray telescopes like *Chandra X-ray Observatory* (CXO), X-ray astronomy has entered a new era. The study of X-ray sources help us answer questions like the existence of Supermassive Black holes, formation and dynamics of galaxies (Grimm et al. 2003), formation evolution and stability of Globular Clusters (Pooley et al. 2016), and so on. All-sky surveys with such advanced telescopes contain a wealth of knowledge with the properties available for millions of sources. A significant fraction of these sources remain unidentified. *Chandra X-ray Observatory* is far Superior with its capabilities compared to its predecessors. *Chandra*'s angular resolution is  $\leq 1''$ , which is 10 times of *ROSAT* and about 40 times that of the *XMM-NEWTON*. Due to the incredible sub-arcsec spatial resolution of *Chandra*, the spatial source population density in the *Chandra* Source Catalogue (CSC) 2.0 is the highest among other X-ray telescopes and their survey catalogues. This allows the CSC 2.0 to be used to identify and study the population distribution of sources in dense field like, globular clusters and can point sources in external galaxies. Heinke et al. (2005) identified 300 X-ray sources within  $2.79''$  radius of 47-Tucane globular cluster. With *Chandra*, they could identify sources as faint as  $L_X < 8 \times 10^{29} \text{ ergs s}^{-1}$ .

The second version of Chandra Source catalogue (CSC 2.0) (Evans

et al. 2019) contains properties of 317,000 sources from the observations till the end of 2014 and a total sky coverage of  $558.65 \text{ deg}^2$  (Evans 2020). Identification of this huge population of surreptitiously detected sources becomes important for any population study. When the number of sources is small, they can be identified using manual methods using their photometric color-color diagram (Schulz et al. 1989; Pollo et al. 2010), timing and spectral properties. These manual methods, due to their time complexity are infeasible when the number of sources is very high. The sheer volume of data generated by the all-sky survey of these telescopes necessitates taking an automated approach for 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 like identifying sources based on the color-color diagram clustering, which are limited to 3 dimension at maximum, 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 fairly recent. Zhang et al. (2021) used Random Forest and LogiBoost to classify the sources in *XMM-newton* 4XMM-DR9 multiwavelength properties from *GAIA*, *WISE*, and *2MASS*. Farrell et al. (2015) did source classi-

\* E-mail: kumaranshivam57@gmail.com

fication of variable sources in the Third *XMM-Newton* Serendipitous Source Catalogue (3XMM) using a Random Forest classifier with timing properties. 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.

In the case of *Chandra*, to the best of our knowledge, no such work of automated classification has been published so far. The catalogue is highly rich as compared to other X-ray catalogues and offers marvellous opportunity for serendipitous discovery of objects of known classes as well as new weird exotic objects ([Martinez Galarza et al. 2019](#)). An automated classifier would open up the possibility of large scale population study of X-ray sources, or in globular clusters ([Heinke et al. 2005](#)), identify point X-ray sources in the external galaxies which would help to study their formation and evolution ([Grimm et al. 2003](#)). The automated classification will lead to identification of thousands of new sources belonging to the known classes, enriching our knowledge base.

In this paper we discuss the development of an automated classifier based on ML algorithms for the point source in CSC 2.0. The classifier primarily use the features available in the CSC 2.0 which are flux in five different bands of *Chandra*’s ACIS instrument, hardness ratio and variability properties. In addition to the X-ray features, the source’s identification can be improved with the use of features available in other wavelengths. We obtain multiwavelength features from 2MASS, *WISE*, *Gaia*-EDR3, *Spitzer*’s 24 $\mu$  MIPS, SDSS and *GALEX*. We explore decision tree based supervised machine learning classification algorithms. Light Gradient Boosted Machine give the best classification performance. In the section 2 we describe the details of the data source and data collection and processing procedure and also describe the method for identifying the training set. In section 3, we describe the classifier models that we have explored, the methodology for selection and validation of the classifier. In section 4, we give the result of model validation and performance evolution. Finally we apply our trained model to the set of unidentified variable sources in CSC 2.0, the application and discussion is given in the section 4.

## 2 THE DATA

The *Chandra X-ray Observatory* (CXO), launched in 1993, is one of the most advanced X-ray observatory in Earth’s orbit. *Chandra* have two focal plane instruments : Advanced CCD Imaging Spectrometer (ACIS) and High Resolution Camera (HRC). ACIS instruments takes observation in the science energy bands : broadband(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, hard (h): 2.0-7.0 keV. HRC instrument take observation in 0.1-10 keV energy band and is designated as ‘W’ band. With the observation from ACIS and HRC till the end of 2014, second version of Chandra Source Catalogue (CSC 2.0) ([Evans et al. 2018](#)) was prepared that contains the information of 317,167 sources out of which 296473 are point sources. For our classification target we choose point sources in the CSC 2.0, selected by the parameter *extent\_flag* == 0 in the catalog. 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* (refer table 1 for description of these flags) After

**Table 1.** Quality flags for sources available in Chandra Source Catalogue - 2.0.

flag code	flag description
<i>pileup_flag</i>	Pileup sources
<i>sat_src_flag</i>	saturated source
<i>conf_flag</i>	Confused source
<i>streak_src_flag</i>	source on CCD streak

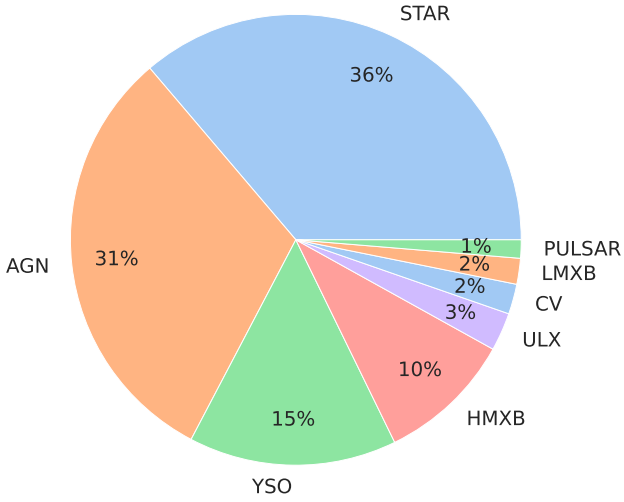
removing the sources for which the values are 1 for the above flags, we get a list of 277717 point sources. We obtain the flux values (in b, u, m, s, and h bands), the variability properties (both inter and intra-observation variability) and hardness ratio (refer to table 2 for these properties) We use CIAO 4.14 to download the data from CSC using Astronomical Data Query Language(ADQL).

The *Wide-field Infrared Survey Explorer* (*WISE*) ([Wright et al. 2010](#)) is an all sky in the infra-red survey in the wavelength bands of 3.4, 4.6, 12, and 22  $\mu$ m named W1, W2, W3 and W4 bands respectively. With *WISE*, a catalogue of 747,634,026 sources with limiting sensitivities W1<17.1, W2<15.7, W3<11.5 and W4<7.7 magnitude. In AllWISE catalog, the SNR=5 is achieved for flux at 54, 71, 730 and 5000 mJy (16.9, 16.0, 11.5 and 8.0 mag) in W1, W2, W3 and W4. In our work we use W1, W2, W3 and W4 magnitude from AllWISE catalog.

The *Gaia* is an optical telescope launched and operated by European Space Agency ([Gaia Collaboration et al. 2016](#)). 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<sub>BP</sub>* and red: *G<sub>RP</sub>* passbands. The limiting magnitude is  $G \approx 21$  and the upper magnitude limit is  $G \approx 3$ . We have used Gaia-EDR3 *G*, *G<sub>BP</sub>* and *G<sub>RP</sub>* magnitudes.

We use NASA/IPAC Extragalactic Database (NED) to obtain multi wavelength information for the CSC sources of our interest. With the November 2021 release, NED integrated the Chandra Source Catalogue and using its cross-match algorithm Match Expert (MatchEx) ([Ogle et al. 2015](#)), 80% of the sources in CSC were associated with already existing NED sources and 20% were added as new source. We use CSC 2.0 names as the object identifier in NED to obtain the multiwavelength property with identifier based query using *astroquery* package. For 648 objects, NED server responded with error message and the data could not be obtained. Nevertheless, this amount to only  $\approx 0.2\%$  of the total number of sources. Finally we obtain multiwavelength data for 277069 sources from NED. Using the NED we obtain the Infra-red data from 2 Micron All Sky Survey (2MASS) catalog, *Wide-field Infrared Survey Explorer* (*WISE*)’s ALLWISE catalogue and *SPITZER*’s Multi-Band Imaging Photometer (MIPS). We obtain Optical/UV data from Sloan Digital Sky Survey (SDSS) and *GALEX* All-Sky Survey Catalogue (GASC). *Gaia* data is not available on the NED server, so we obtain *Gaia*’s Optical magnitudes from from *Gaia* EDR-3 catalog’s server. We find the association with *Gaia* using CDS X-match positional cross match service ([Boch et al. 2014](#)) with cross matching criteria to be within  $3\sigma$  of the positional error of CSC 2.0 and *Gaia* EDR-3 error circle. For the 277069 CSC 2.0 sources we could find an association for 60% sources in *Gaia*-EDR3, 55% in 2MASS, 43% for *Spitzer*’s 24  $\mu$  MIPS, 41% for *WISE*, 24% for SDSS and 17% for *GALEX*.

The multiwavelength features that we are using for these catalogues, their names and description is given in the table 2. Other than the features obtained from the catalogues, we have used the magnitude available in different bands to compute the color and use them as the feature. To identify the which colors to be calculated, we



**Figure 1.** This pie chart shows the percentage of sources in the training set belonging to each class. This pie chart clearly highlights the imbalance in the population distribution across the class

have used the online multiwavelength visualization tool developed by Yang et al. (2021).

## 2.1 Identification of training set

For 277069 point sources in CSC 2.0, we populate the data-table with 41 MW features (refer to the table 2 for more details of the features) with prime data source CSC 2.0 itself and multi wavelength data from GAIA-EDR3, 2MASS, AllWISE, SDSS, GALEX and MIPS. We design the classifier to identify the class belonging to one of the classes in Active Galactic Nuclei (AGN), X-ray emitting stars (STARS), Young Stellar Objects (YSO), Low Mass X-ray binaries (LMXB), High Mass X-ray binaries (HMXB), Cataclysmic Variables (CV), pulsars and Ultra Luminous X-ray sources (ULX). First we prepare a list of all the already identified sources belonging to these classes and then cross match their published position with CSC 2.0 sources. The catalogue that we use to identify known sources and their position alongwith the catalogue reference is given in table 4.

Using the coordinates of the known source, we cross match these with all the 'good' sources in CSC 2.0. We select a cross match radius of  $1''$ . We use ASTROPY, which is a PYTHON package to perform cross-matching. In case of more than one cross matching sources within  $1''$ , we consider the source with the least angular separation of the either. Using this we identified a total of 7703 sources. The class-wise number of identified sources are given in table 3.

The pie chart in the figure 1 shows the percentage of population of the sources in each class. In our training set we have a large fraction of AGNs, YSOs and Stars which comprises of a total of about 80 percent of the entire training set. The classes LMXB, CV, pulsar and ULX are minority class with population only 1-3 percent of the training set.

## 3 METHODOLOGY

### 3.1 Classifier Models

We prepare the multi-wavelength dataset and select a set of already identified sources. We use this set to train a supervised machine learning model, where the model learn pattern from 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 tested the models available with their default parameters available in Scikit-Learn (Pedregosa et al. 2011) Python library. From the Scikit-Learn library we identified that decision tree based models: Random Forest (RF) and Gradient Boosted Decision Tree (GBDT) gave the best result compared to other models. We tried Light Gradient Boosted Machine (LightGBM) (Ke et al. 2017), which is an advanced development over GBDT.

#### 3.1.1 Random Forest

Random Forest (RF) (Breiman 2001) is an ensemble of decision tree. In the ensemble, each decision tree is built from a randomly selected bootstrapped sample from the training set. Each tree thus built is unique in nature and is independent from each other and acts as a parallel weak learners. For a given sample, each tree votes for it belonging to one of the classes. For a given sample the fraction of trees out of the entire ensemble, voting for a particular class is treated as the class membership probability.

#### 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 trees are built independently, in GB the trees are built sequentially based on the error of previous tree. With a given loss function the loss at each newly constructed tree is built. The gradient of this loss function at  $m - 1^{th}$  tree is used to construct a new tree. This new tree is combined to the previous trees after multiplying it with a weight factor called Learning rate  $\eta$ , which generally varies from 0 to 1. Intuitively new tree is built to minimize the error from the previous tree. For classification, generally categorical cross entropy is chosen as the loss function. The main advantage of GB over RF, is that in each newly constructed tree, the tree function uses the loss from previous tree. Intuitively, each new tree tries hard to classify previously incorrectly classified sources. Compared to Random Forest, Gradient Boosted trees can learn more complex decision boundaries.

#### 3.1.3 Light Gradient Boosted Machine

Light Gradient Boosted Machine (LightGBM) was developed by Microsoft 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, in LightGBM first bins the continuous values and then use these binned value to create the decision tree. To make the learning using Gradient Boosting more efficient, LightGBM implements two novel techniques 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 Exclusive feature bundling LightGBM bundles the mutually exclusive features (the features that rarely takes zero simultaneously) to reduce the feature space. Another major feature of LightGBM is that it can handle missing values. It

**Table 2.** Multi-wavelength features and their description. Note - the 'Color' features are not from any specific catalogue and are computed using available magnitude values

Feature Source	Feature Name	Feature Description
CSC 2.0	gal_l2	Galactic longitude
	gal_b2	Galactic Latitude
	h-csc	Flux in ACIS hard (h) band (2.0-7.0 keV)
	m-csc	Flux in ACIS medium (m) band (1.2-2.0 keV)
	s-csc	Flux in ACIS soft (s) band (0.5-1.2 keV)
	u-csc	Flux in ACIS ultrasoft (u) band (0.2-0.5 keV)
	b-csc	Flux in ACIS broad (b) band (0.5-7.0 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
GAIA-EDR3	kp_intra_prob_b	Intra-observation Kupier's test variability probability in b band
	var_intra_index_b	Intra-observation variability index
	G	Gaia Green (G) pass-band magnitude
	Bp	Gaia Blue (G_BP) pass-band magnitude
	Rp	Gaia Red (G_RP) pass-band magnitude
	FUV	Magnitude in GALEX FUV band
	NUV	magnitude in GALEX NUV band
	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	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	Ks-band (2.159 micron) band magnitude
Computed Color	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

**Table 3.** Number of sources in the training set (the sources from the published identified source list which cross-matched with CSC sources within a radius of 1 arcsec) and the unidentified sources

Class	Number of sources
AGN	2395
STAR	2790
YSO	1149
HMXB	748
LMXB	143
ULX	211
CV	166
PULSAR	101
<b>Total training set</b>	<b>7703</b>
Unidentified Sources	269366
<b>Total</b>	<b>277069</b>

is a very significant feature for use as we have very high amount of missing values in our dataset. It uses Block Propagation method (Josse et al. 2019). Nodes are spitted using only available feature values and the missing values are sent to the side which minimizes the final error.

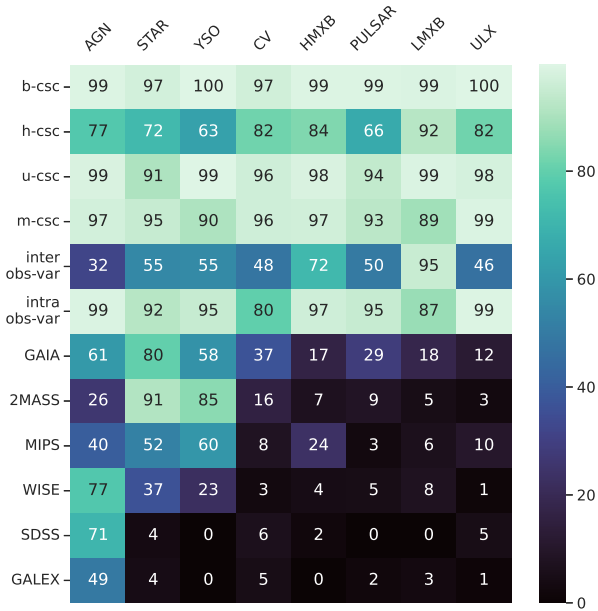
### 3.2 Missing value imputation

In our data-table, we compile features from different multi-wavelength catalogues. Due to difference in coverage in these catalogues and the difference in limiting sensitivity, some objects may not be observed in some of the catalogues. For example the SDSS survey coverage is only limited to northern hemisphere. Due to difference in the intrinsic luminosity of the source combined with the difference in limiting sensitivity across different wavelengths, we have missing values in the data table where associations could not be found. For example, X-ray binaries in quiescent stage have lower luminosity in Optical-UV and IR but are prominent in X-ray. Across difference



**Table 4.** Table showing the source of various catalog, their details and the corresponding reference for class identification of training sources.

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 Catalog	Shaw Greening et al. (2009)
ULX	ULXRBCAT		Liu & Mirabel (2005)
CV	The Open CV Cat.	The Open Cataclysmic Variable Catalogue	Jackim et al. (2020)
	ATNF		Manchester et al. (2005)
PULSAR	FERMI LAT (4FGL)	Fermi LAT Second Catalogue of Gamma-Ray Pulsars (2PC)	Abdo et al. (2013)

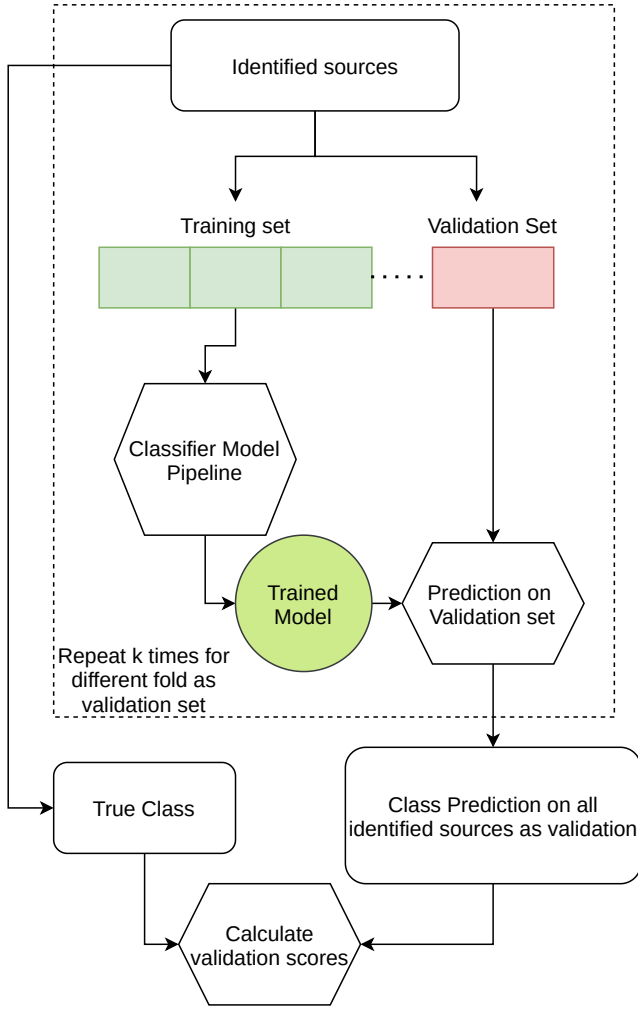
**Figure 2.** Color-map plot showing the percentage of availability for different features group. For better visual inference, the percentage value is color coded.

features within X-ray domain, based on the variability timescales of the objects, the variability features are not available for some of the objects. of the survey instruments, objects may not be available in all the catalogue in the Missing values in the data-table is inevitable in this case. 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 case of RF and GBDT models, we select to impute the missing values using column

mode values. However the imputation creates artificial skew in the dataset and makes the result less reliable. Also in some cases, these missing values itself may be an 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 tried to design the classifier with Light Gradient Boosted Machine. This model can work with missing values in the input feature.

### 3.3 Class imbalance problem

From the table 3 and the figure 1, it is obvious that there is a vast imbalance in the number of training sources, with majority classes being AGN, YSO and Stars and minority classes are LMXB, ULX, CV and pulsar. AGN and stars are 30-35% higher in number than that of LMXB, ULX, CV and pulsars. 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 have used Synthetic Minority Oversampling Technique(SMOTE) (Chawla et al. 2011). In the feature space, linear interpolation between k-neighbouring points (which represents a source in feature space) are done. With this linear interpolation synthetic sources are sampled. Using this technique each class is sampled in such a way as in final training sample, the number of becomes equal to the number of sources in the most populous class. To keep our result insensitive to the oversampling, SMOTE is done only on the training set and not on validation set. SMOTE is done only for the model RF and GBDT and not for LightGBM. In LightGBM we are working with missing values and SMOTE can not be performed with missing values. In LightGBM, we have used class weight technique. In this method, higher weightage is assigned to the samples belonging to the minority class in the calculation of loss function while training. In essence In the loss function, we make sure that equal contribution comes from each class. Any class balancing methods are not capable of reproducing the case of equal class dis-



**Figure 3.** Flowchart showing the cumulative cross validation algorithm. The components inside dashed box represents one fold of validation, model is trained on  $k$  such folds and then the results are accumulated. The final scores are computed over this accumulated predictions. The algorithm is discussed in detail in section 3.4

tribution. Both the methods, SMOTE and class-weight are as good as the representation of the available sample.

### 3.4 Strategy for model performance validation

We compare the performance of the classifier 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. In this method, we divide the training set into  $k$ -fold, and in each iteration, we train the classifier using  $k-1$  folds and keeping aside  $k^{th}$  fold as validation set. After training we make prediction on the  $k^{th}$  fold (represented by the blocks enclosed within dashed box in the figure 3). Instead of computing the metrics on this fold we accumulate the prediction for each fold when they were the validation set. In this way we are able to compute the validation performance on the entire set and the result we get is more likely to be the one we would expect from the final classifier.

The matrix we use for comparing classifier performance is precision, recall and f1-score. In probabilistic terms, the recall score

is the probability of identifying the samples truly belonging to that particular class, precision score is the probability that predicted class is actually the true class for the sample and f1-score is the harmonic mean of precision and recall:

$$precision_A = \frac{TP}{TP + FP} \quad (1)$$

$$recall_A = \frac{TP}{TP + FN} \quad (2)$$

$$f1\_score = 2 \times \frac{precision \times recall}{precision + recall} \quad (3)$$

where,  $precision_A$  represents the precision score for class  $A$  and so on,  $TP$  represents the number of predictions which are predicted as class  $A$  and actually belong to class  $A$ ,  $FP$  is the number of samples for which prediction are of class  $A$  but actually belong to some other class and  $FN$  represents the number of samples which are predicted as a class other than  $A$  but actually belongs to the class  $A$ .

We have also used Mathew's correlation coefficient (MCC), first introduced by Matthews (1975) which is supposed to be a better choice in case of imbalanced dataset (Boughorbel et al. 2017). MCC is defined as

$$N = TN + TP + FN + FP \quad (4)$$

$$S = \frac{TP + FN}{N}, P = \frac{TP + FP}{N} \quad (5)$$

$$MCC = \frac{TP/N - S \times P}{\sqrt{PS(1-S)(1-P)}} \quad (6)$$

## 4 RESULT AND DISCUSSION

We trained the classifier models, LightGBM, RF and GBDT and evaluated the performance using the cumulative k-fold cross validation. We implemented the models using Sci-kit LEARN package. We compared the classifier performance model using CCV method. For the final selected model, we tuned the hyperparameter using FLAML(A Fast Library for Automated Machine Learning & Tuning) by Wang et al. (2019). To compare the model performance, using CCV method, we computed the precision, recall and f1 score for each classes.

The precision, recall and f1 score for 20-fold cumulative cross validation is given in table 5. The performance for Majority classes AGN, YSO and Star with f1 score > 94% are significantly higher than compare to CV, ULX and pulsar with f1 score about 40 to 60%. Performance is moderately good for LMXB and HMXB. Across the model the same trend is seen. When the models are compared, LightGBM performs the best for each classes. RF and GBDT have similar performance. For LightGBM, the performance is marginally higher than RF for AGN, STAR and YSO, but for LMXB,HMXB and ULX the score is significantly higher. For minority classes, there is a large difference between precision and recall for RF and GBDT models, whereas these two scores are comparable for LightGBM. For example for pulsars, RF precision score is 46% whereas recall score is only 22%, which means that the model is being too careful to predict this class and in doing so it misses about 80% of true pulsars. This factor is well balanced in the case of LightGBM, with precision of 42% and recall as 45%. At the classes' precision and recall score are well balanced for LightGBM compared to other two models, giving the highest f1-score.

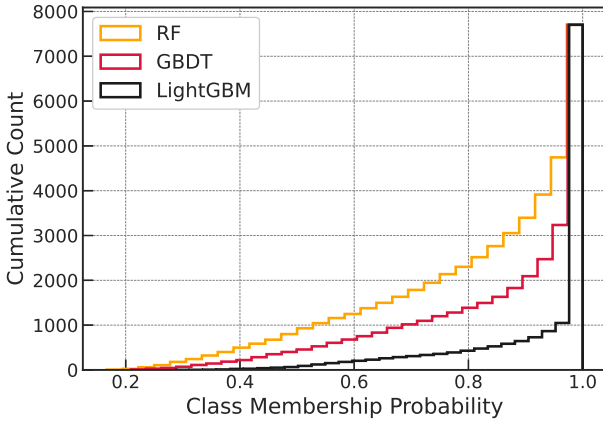
The table 6 shows the MCC score and the macro average of precision, recall and f1-score for RF, GBDT and LightGBM classifier. LightGBM have the highest accuracy score of 93.1% whereas for RF

**Table 5.** Precision, recall and f1 score for different classes for Random Forest(RF), Gradient Boosted Decision Tree(GBDT) and Light Gradient Boosted Machine (LightGBM) models. The scores are calculated by 20-fold cumulative cross validation.

Score→ class↓ Model→	RF	Precision GBDT	LightGBM	RF	Recall GBDT	LightGBM	RF	f1-score GBDT	LightGBM
AGN	94.2±0.2	95.2±0.2	96.8±0.1	96.8±0.2	96.9±0.1	97.6±0.2	95.5±0.1	96.1±0.1	97.2±0.1
STAR	93.9±0.2	93.8±0.2	96.1±0.1	95.4±0.2	95.2±0.1	95.7±0.1	94.6±0.1	94.5±0.1	95.9±0.1
YSO	90.2±0.2	90.7±0.2	92.8±0.2	94.5±0.2	95.4±0.3	95.4±0.3	92.2±0.1	93.00.2	94.1±0.2
HMXB	80.4±0.3	89.1±0.7	91.8±0.2	85.5±0.2	90.0±0.2	90.7±0.7	82.9±0.2	89.6±0.4	91.2±0.4
LMXB	98.7±0.5	92.5±1.8	94.8±1.3	74.5±0.3	79.2±0.3	80.7±0.8	84.9±0.3	85.3±0.8	87.2±0.6
ULX	69.7±0.9	75.9±1.3	71.7±1.2	51.4±1.0	59.3±1.6	71.2±1.1	59.1±0.9	66.6±1.0	71.5±1.0
CV	67.0±2.0	63.6±1.2	60.6±1.2	41.0±0.9	48.9±1.1	54.5±1.5	50.8±1.1	55.3±1.1	57.4±1.3
PULSAR	45.9±2.4	38.56±1.8	42.1±1.9	22.0±1.12	20.8±0.9	45.3±1.9	29.7±1.4	27.0±0.8	43.7±1.9

**Table 6.** Comparison of Accuracy, Precision (average), recall (average) and f1-score (average) for Random Forest, Gradient Boosted Decision Tree and LightGBM classifier models

Score	RF	Model GBDT	LightGBM
Accuracy	90.0 ± 0.2	89.1 ± 0.2	93.3 ± 0.12
Precision	73.2 ± 0.4	70.8 ± 0.6	80.8 ± 0.38
Recall	76.2 ± 0.5	79.4 ± 0.3	78.9 ± 0.33
f1 score	74.5 ± 0.4	73.4 ± 0.4	79.8 ± 0.34
MCC	0.87 ± 0.0	0.87 ± 0.003	0.91 ± 0.002

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

and GBDT have accuracy about 89%. All the scores are highest for the LightGBM classifier.

Apart from the good accuracy score, the ideal classifier should be capable of prediction the class membership with high confidence. The figure 4 shows the cumulative histogram of the class membership probabilities 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 in the plot shows the number of sources with CMP below the value given on X-axis. The more steeper the plot towards 1.0 means the higher the predicted CMPs, the more confident is the model. From the figure the LightGBM is the most confident classifier model among the three with only about 500 out of 7703 sources below CMP of 0.8, with RF being the least with more than

**Table 7.** Class wise precision, recall and f1 score for CCV (with mean and standard deviation calculated for 5 CCVs) using LightGBM model. The validation is done for different set of samples indicated in first column: **All features** is the sample where all the 41 features in the table 2 are used; **no-IR** : the sample with IR features (W1, W2, W3, J, H, K and colors involving these features) removed; **no-optical-uv**: the sample with optical/UV feature (G, Bp, Rp, FUV, NUV and colors involving these features) removed; **no-MW**: the sample with no optical, UV and IR features and only retaining the X-ray and galactic coordinates.

class	type	precision	recall	f1 score
AGN	all-features	96.8±0.1	97.6±0.2	97.2±0.1
	no-MW	91.1±0.1	95.1±0.1	93.0±0.1
STAR	all-features	96.1±0.1	95.7±0.1	95.9±0.1
	no-MW	89.2±0.2	88.2±0.1	88.7±0.1
YSO	all-features	92.8±0.2	95.4±0.3	94.1±0.2
	no-MW	82.9±0.1	89.4±0.3	86.0±0.2
HMXB	all-features	91.8±0.2	90.7±0.7	91.2±0.4
	no-MW	92.3±0.5	90.1±0.5	91.2±0.4
LMXB	all-features	94.8±1.3	80.7±0.8	87.2±0.6
	no-MW	94.5±1.0	81.7±0.5	87.6±0.6
ULX	all-features	71.7±1.2	71.2±1.1	71.5±1.1
	no-MW	61.9±1.3	43.6±1.3	51.2±1.4
CV	all-features	60.6±1.2	54.5±1.5	57.4±1.3
	no-MW	56.1±2.8	45.1±1.0	50.0±1.7
PULSAR	all-features	42.1±1.9	45.3±1.92	43.7±1.9
	no-MW	28.2±1.9	18.6±0.7	22.4±0.9

2000 sources below CMP of 0.8. From the validation comparison, LightGBM is a clear classifier model of choice.

To study the importance of different features we used feature elimination technique. Apart from the one base sample, in which we have included all 41 features mentioned in table 2, we created three different samples, with one set of feature removed in each of them. The LightGBM classifier model is validated with 5 times 20-fold cumulative cross validation. Then we compared the model performance with the base sample score to see the class-wise dependency of the feature.

- **All features:** All the 41 features in table 2 included.
- **no-MW** : Removed all IR, Optical and UV features including colors. Total number of features used - 14

The table 7 shows the mean and standard deviation of the precision, recall and f1 score over 5 times of the 20-fold cumulative cross validation. For each class, the table shows the scores for the two cases: one with X-ray features combined with MW 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 pulsars and least for the AGN. For pulsars, the f1 score drops from 47% to 22%. Considering recall score, with all features, 45% of the pulsars could be retrieved however with only X-ray features only 18% of the pulsars could be retrieved. For AGN there is only a marginal drop in the score. 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 remarkably higher f1 score of 93%. Stars too follows a similar trend but for stars, the dependency on Optical/UV/IR features is higher as there is drop of 8% in f1 score with MW features removal. YSO's can be identified with only X-ray features with 86% f1 score, which improves to 93% when Optical/UV, IR features are also used. HMXB and LMXB on the other hand are insensitive to Optical/UV and IR features. With no drop in performance on any metric with drop in MW features. This can be attributed to the fact that for LMXB and HMXB except for MIPS, the values are mostly unavailable in SDSS, 2MASS, WISE and GALEX. X-ray features are the most important for identification of these two classes. For the ULX when all the MW features are removed, the drop is about 20%. In the absence of MW features, ULX are more likely to be classified as AGNs, which results in slight drop in AGN score but due to very small population, this translates to a very high drop in ULX. For CV both Optical/UV and IR features are equally important but results in a drop of only 7% when only X-ray features are used. For CV X-ray features are important with MW features not able to bring a significant improvements. The case of pulsar is interesting. The precision score drops by 10% and 5% for IR and Optical/UV features respectively. Without IR feature, the recall score drops by about 20% but without Optical/UV feature this drop is only about 5%. Which means that IR feature more of the true pulsars are identified correctly. IR have higher importance for pulsar than Optical/UV features. Also with inclusion of MW features pulsars shows the most significant improvement in identification across all the classes, with f1 score 20% more than the case on only X-ray features. The figure 5 shows the confusion matrix for the two cases: one where all the features are used and the one with only X-ray features. The element of the matrix shows the fraction of sources which truly belong to the class given on y-axis being classified as that of belonging to the class given on x-axis. The diagonal element essentially are the recall score for the individual classes. Looking at the 'ULX' row, with MW features only 9.5% of ULX were identified as AGN, but with only X-ray features 41% of ULXs are identified as AGNs. Using MW features, the classifier is better able to separate ULX from AGN. Without MW features, pulsars are most likely being identified as Star as after removing X-ray features fraction of pulsars identified as Star grows from 11% to 40%. From the confusion matrix, it is evident that without MW features, the network becomes more biased towards AGN, YSO and Stars which are majority classes, as the fraction of ULX, CV and pulsar being classified as AGN, Star and YSO increases.

From the results discussed above, we choose LightGBM as the final classifier with all the 41 features in the table 2 for training. We then train it using all the 7703 sources in the training set (see table 3). For a given source the model assigns a class membership probabilities(CMP), which says the probability of it belonging to each class. For the source the class is chosen based on the maximum of these CMP. The trained LightGBM model is then applied to all the 269366 unidentified sources. The classifier assigns a CMP in the range 0 to 1 for all the given source. This CMP is a continuous variable to find out the class wise distribution of the CMP we plot the probability density function which is given in the figure 6. The probability of getting a CMP= $x$  between the values  $a$  and  $b$  is given

**Table 8.** The number of sources identified by the classifier, with a selection scheme based on the maximum class membership probability in the third column and the number of sources identified above 3-sigma confidence (class membership probability  $>0.997$ ) is shown in the fourth column and above  $4\sigma$  confidence (class membership probability  $>0.9999$ ). Second columns shows the most probable value for class membership probability for each class.

Class	Mode(CMP)	Number of sources		
		all	$CMP > 3\sigma$	$CMP > 4\sigma$
AGN	0.973	114642	32600	8574
STAR	0.967	63967	16148	5166
YSO	0.967	40524	5184	208
HMXB	0.943	8321	439	46
LMXB	0.973	1688	197	71
ULX	0.528	6083	50	0
CV	0.502	10999	89	1
PULSAR	0.552	23142	63	0
<b>Total</b>		<b>269366</b>	<b>54770</b>	<b>14066</b>

by the area under the curve in the probability density plot and is described by the integral given in the equation 7

$$P(a < x < b) = \int_a^b f(x)dx \quad (7)$$

. The curve in the pdf represents the function  $f(x)$ . This value of  $f(x)$  is calculated from the CMP histogram with  $N$  bins,  $\Delta x$  bin-size and  $n_i$  counts in  $i^{th}$  bin using the following equation :

$$f(x_i) = \frac{n_i}{\sum_0^N n_i \times \Delta x} \quad (8)$$

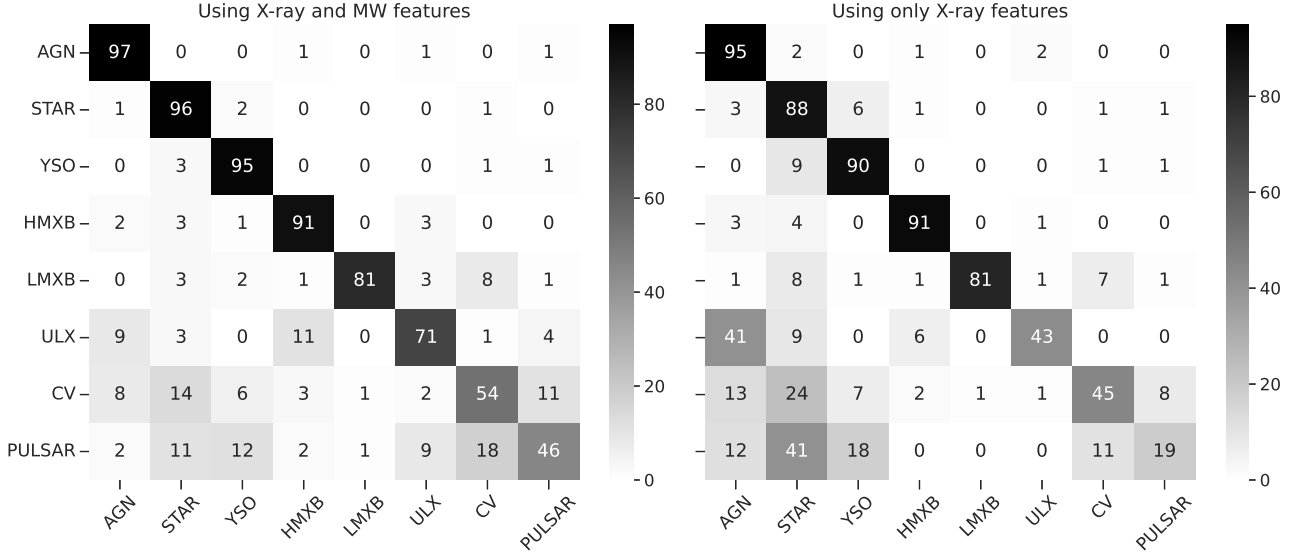
where Figure 6 shows the probability density function (PDF) of the class-wise CMP we got after applying the trained model to the unidentified sources. The probability density function is calculated using Kernel density estimation (KDE) method. The more narrower the curve towards 1.0, the better is the predicted confidence. The figure also shows the peak of the PDF using red dashed lines, which is the most probable CMP value for the given class. For AGN, Star and YSO the PDF shows a sharp peak close to 1.0. The plot shows that the LightGBM model is able to predict most of the AGN, YSO, STAR, HMXB and LMXB more with a very high confidence. For pulsar, CV and ULX getting CMP  $>0.9$  is very less likely than getting CMP of  $\sim 0.5$ . The pdf also helps in calculating the most probable value for the CMP. For example, from all the sources AGNs are most likely to be identified with a probability close to 0.97, whereas this value for CV is close to 0.5. For reporting the final classification instead of assigning the class based on the maximum CMP, we can set a higher threshold to select only those sources for which the CMP is above that threshold.

The table 8 shows the most probable values of CMP for each class in the first column. The number of sources when we assign classes for all the sources based on the maximum CMP across class for each source is given in the second column. We selected a probability confidence threshold of  $3\sigma$ , which is  $CMP > 0.997$  to get a set of confidently classified sources, the numbers of which is shown in the last column of the table 8.

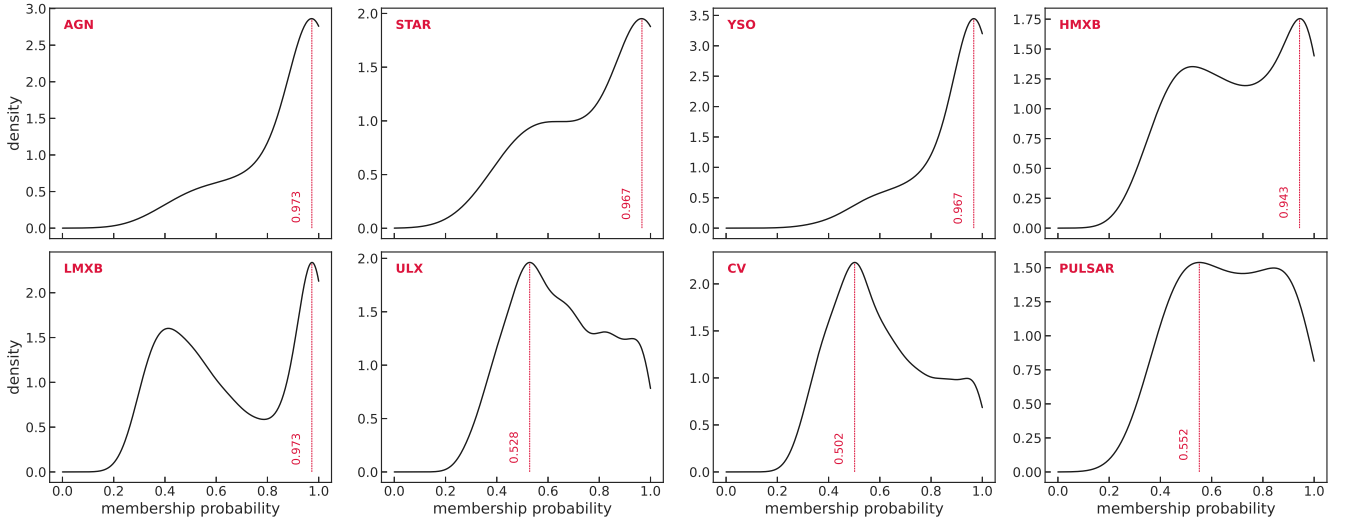
## 5 SUMMARY AND CONCLUSIONS

Many objects of the *Chandra* Source Catalogue CSC 2.0, which contains  $\approx 317,000$  sources, including  $\approx 277,000$  point sources, were unidentified. In this work, we implement the decision tree based





**Figure 5.** Confusion matrix for the two sample set: with X-ray and other MW features(optical,IR,UV refer to table 2 for details) on the left and With only X-ray features on the right. The number in these matrix are the percentage of sources belonging to the class on vertical axis being classified as that of the class given on the horizontal axis.



**Figure 6.** Probability density function (pdf) of the distribution of class membership probabilities of all the unidentified sources predicted from our model for different classes (marked on the plots). The pdf is computed using kernel density estimation method. the probability of occurrence of a 'membership probability' in a range can be computed as the area under the curve in the given range (see equation 7 for details). The most probable values of the class membership probability is shown in the plot with vertical dashed lines on each class's plot.

classifier Light Gradient Boosted Machine to identify the CSC 2.0 objects in the classes of AGN, YSO, star, HMXB, LMXB, ULX, CV, and pulsar. Our method gives class membership probabilities of these sources. For the classification, we use X-ray properties from *Chandra*, optical properties from *Gaia*, *SDSS* and *GALEX*, and Infrared properties from *2MASS*, *WISE* and *MIPS*. Our classifier has the classification accuracy of 93%, precision score of 81%, recall score of 79%, f1 score of 80% and Mathew's Correlation coefficient of 0.91. Thus, this can be a reliable and promising classifier for other catalogues as well. With this classifier, we identify 54770 new point sources. Among them, there are 32600 AGNs, 16148 stars, 5184 YSOs, 197 LMXBs, 439 HMXBs, 50 ULXs, 89 CVs and 63 pulsars

with a confidence of more than  $3\sigma$ . At a higher confidence of more than  $4\sigma$ , we get 8574 AGNs, 5166 Stars, 208 YSOs, 46 HMXBs, 71 LMXBs and 1 CV but not ULX 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.

## ACKNOWLEDGEMENTS

This research has used the Chandra Data Archive data and the Chandra Source Catalog, 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 color-color properties to be used.

## DATA AVAILABILITY

The inclusion of a Data Availability Statement is a requirement for articles published in MNRAS. Data Availability Statements provide a standardised format for readers to understand the availability of data underlying the research results described in the article. The statement may refer to original data generated in the course of the study or to third-party data analysed in the article. The statement should describe and provide means of access, where possible, by linking to the data or providing the required accession numbers for the relevant databases or DOIs.

## REFERENCES

- Abdo A. A., et al., 2013, *ApJS*, **208**, 17
- 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
- Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P., 2011, *arXiv e-prints*, p. [arXiv:1106.1813](https://arxiv.org/abs/1106.1813)
- Ćiprijanović A., et al., 2021, *MNRAS*, **506**, 677
- Ebisawa K., Bourban G., Bodaghee A., Mowlavi N., 2003, *Astronomy & Astrophysics*, **411**, L59
- Evans I., 2020, *Chandra News*, **28**, 6
- Evans I. N., et al., 2018, in *American Astronomical Society Meeting Abstracts* #231. p. 238.01
- Evans I. N., et al., 2019, in *AAS/High Energy Astrophysics Division*. p. 114.01
- 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
- Gaia Collaboration et al., 2016, *A&A*, **595**, A1
- Grimm H.-J., Gilfanov M., Sunyaev R., 2003, *Chinese Journal of Astronomy and Astrophysics Supplement*, **3**, 257
- Haakonsen C. B., Rutledge R. E., 2009, *ApJS*, **184**, 138
- Heinke C. O., Grindlay J. E., Edmonds P. D., Cohn H. N., Lugger P. M., Camilo F., Bogdanov S., Freire P. C., 2005, *ApJ*, **625**, 796
- 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
- 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. 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
- 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
- Ogle P. M., Mazzearella 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 and Systems XXIV (ADASS XXIV)*. p. 25 ([arXiv:1503.01184](https://arxiv.org/abs/1503.01184))
- Pedregosa F., et al., 2011, *Journal of Machine Learning Research*, **12**, 2825
- Pietsch W., Freyberg M., Haberl F., 2005, *A&A*, **434**, 483
- Pollo A., Rybka P., Takeuchi T. T., 2010, *A&A*, **514**, A3
- Pooley D., Pooley D. 2016, *MmSAI*, **87**, 547
- Ritter H., Kolb U., 2003, *A&A*, **404**, 301
- Schulz N. S., Hasinger G., Truemper J., 1989, *A&A*, **225**, 48
- 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
- Sturm R., et al., 2013, *Astronomy & Astrophysics*, **558**, A3
- Tang H., Scaife A. M. M., Leahy J. P., 2019, *MNRAS*, **488**, 3358
- Tezhinsky 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](https://arxiv.org/abs/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
- Zhang Y., Zhao Y., Wu X.-B., 2021, *MNRAS*, **503**, 5263
- de Beurs Z. L., Islam N., Gopalan G., Vrtillek S. D., 2022, *arXiv e-prints*, p. [arXiv:2204.00346](https://arxiv.org/abs/2204.00346)

This paper has been typeset from a  $\mathrm{\TeX}/\mathrm{\LaTeX}$  file prepared by the author.