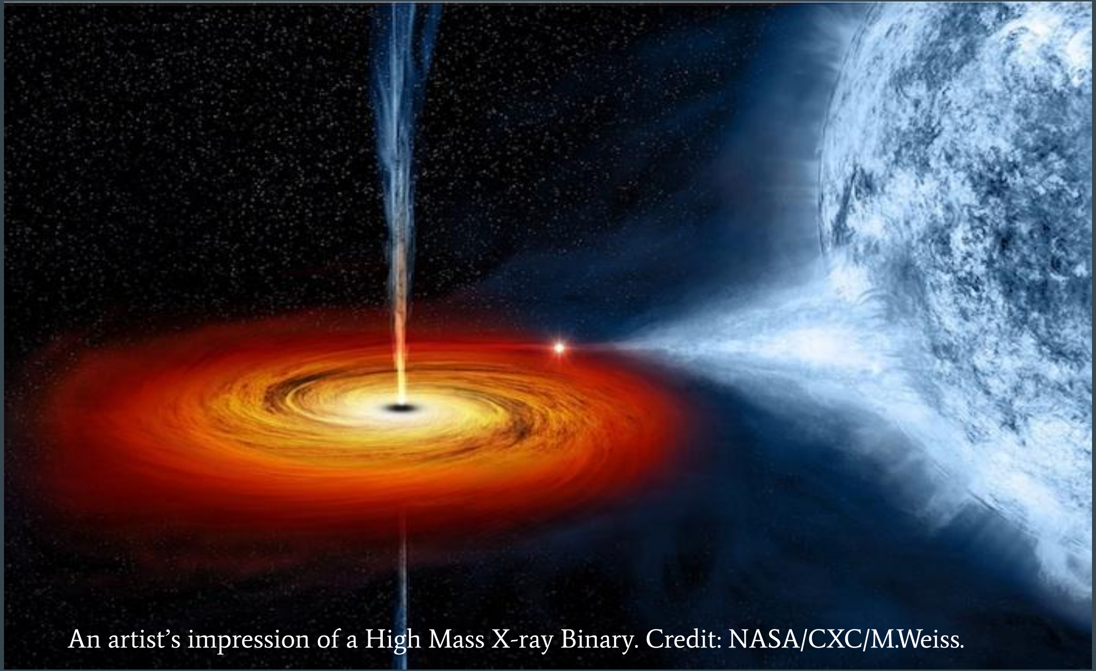
“Identifying New X-Ray Binary Candidates in M31 using Random Forest Classification”

Elias Kyritsis

Journal Club-11/06/2020

Institute of Astrophysics, Crete



An artist’s impression of a High Mass X-ray Binary. Credit: NASA/CXC/M.Weiss.

Identifying New X-ray Binary Candidates in M31 using Random Forest Classification

R. M. Arnason[1]*, P. Barmby[1,2], N. Vulic[1,3,4]

[1] Department of Physics and Astronomy, University of Western Ontario, 1151 Richmond Street, London, ON N6A 3K7, Canada
[2] Institute for Earth and Space Exploration, University of Western Ontario, 1151 Richmond Street, London, ON N6A 3K7, Canada
[3] Laboratory for X-ray Astrophysics, Code 662, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA
[4] Department of Astronomy and Center for Space Science and Technology (CRESST), University of Maryland, College Park, MD 20742-2421, USA

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Identifying X-ray binary (XRB) candidates in nearby galaxies requires distinguishing them from possible contaminants including foreground stars and background active galactic nuclei. This work investigates the use of supervised machine learning algorithms to identify high-probability X-ray binary candidates. Using a catalogue of 943 *Chandra* X-ray sources in the Andromeda galaxy, we trained and tested several classification algorithms using the X-ray properties of 163 sources with previously known types. Amongst the algorithms tested, we find that random forest classifiers give the best performance and work better in a binary classification (XRB/non-XRB) context compared to the use of multiple classes. Evaluating our method by comparing with classifications from visible-light and hard X-ray observations as part of the Panchromatic Hubble Andromeda Treasury, we find compatibility at the 90% level, although we caution that the number of source in common is rather small. The estimated probability that an object is an X-ray binary agrees well between the random forest binary and multiclass approaches and we find that the classifications with the highest confidence are in the X-ray binary class. The most discriminating X-ray bands for classification are the 1.7–2.8, 0.5–1.0, 2.0–4.0, and 2.0–7.0 keV photon flux ratios. Of the 780 unclassified sources in the Andromeda catalogue, we identify 16 new high-probability X-ray binary candidates and tabulate their properties for follow-up.

Key words: X-rays:binaries – X-rays:galaxies – galaxies:individual:Andromeda – techniques:statistical – stars: black holes – stars: neutron

Introduction

Definition of XRBs

X-ray binary system

Compact Object (Accretor) + Companion Star (Donor)

Black Hole

Neutron Star

White Dwarf

Introduction

Classification of XRBs

X-ray binary systems are classified mainly by the mass of companion star

Low Mass X-ray Binaries	High Mass X-ray Binaries
LMXBs	HMXBs
Companion star: $M < 1\ M_{\odot}$	Companion star : $M > 10\ M_{\odot}$
Spectral type : Later than B	Spectral type : O or B
Accretion through a Roche Lobe overflow	Accretion through stellar wind

XRBs can also be categorized by the type of compact object accreting material from the companion star

Introduction

Why should one study the XRBs ?

Excellent labs of extreme physics + Tracers of galaxy properties !

HMXBs

Tracers of current star formation in a galaxy

- XLFs of sources within star-forming galaxies are dominated by contributions of these XRBs.
- In the Galaxy, cluster spatially close to active star-forming complexes

LMXBs

Tracers of past star formation and current stellar density in a galaxy

- Low mass stars comprise the bulk of any stellar population in a galaxy
- Are found in the globular clusters of galaxies due to high stellar densities which enable dynamical encounters

Accurate determination of the XRB number in a population is required !

Introduction

Why should one study the XRBs ?

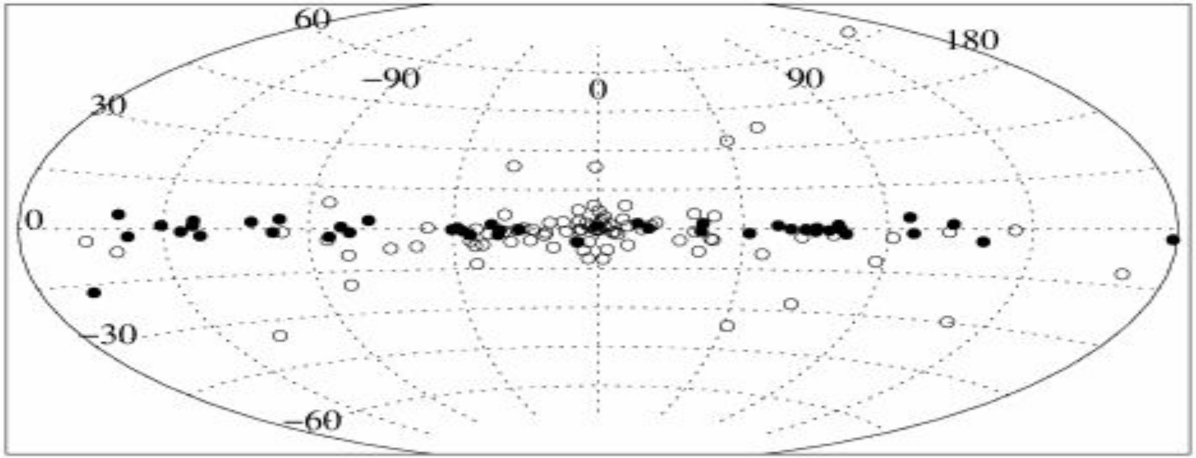


Fig. 1 Distribution of LMXBs (open circles) and HMXBs (filled circles) in the Galaxy. In total 86 LMXBs and 52 HMXBs are shown. Note the significant concentration of HMXBs towards the Galactic Plane and the clustering of LMXBs in the Galactic Bulge.

[Grimm et al. , 2003]

Identifying X-Ray Binaries

XRBs population studies in Nearby galaxies vs Milky way

MW: Suffers from distance uncertainties & Dust + gas in the disk obscure our line of sight

Nearby galaxies: All sources in the same distance & Resolving the structure at a favourable viewing angle without affecting the detection of X-ray source populations (i.e M31)

X-Ray source lists in nearby galaxies contaminated by:

- X-ray active foreground stars in the MW
- Background AGNs
- SNRs

Identification of IR or optical counterparts can solve this problem

BUT

Multiwavelength observations may not be available due to extinction or large distance

Identifying X-Ray Binaries

Solution : Taking advantage of the unique signatures in their X-Ray spectra

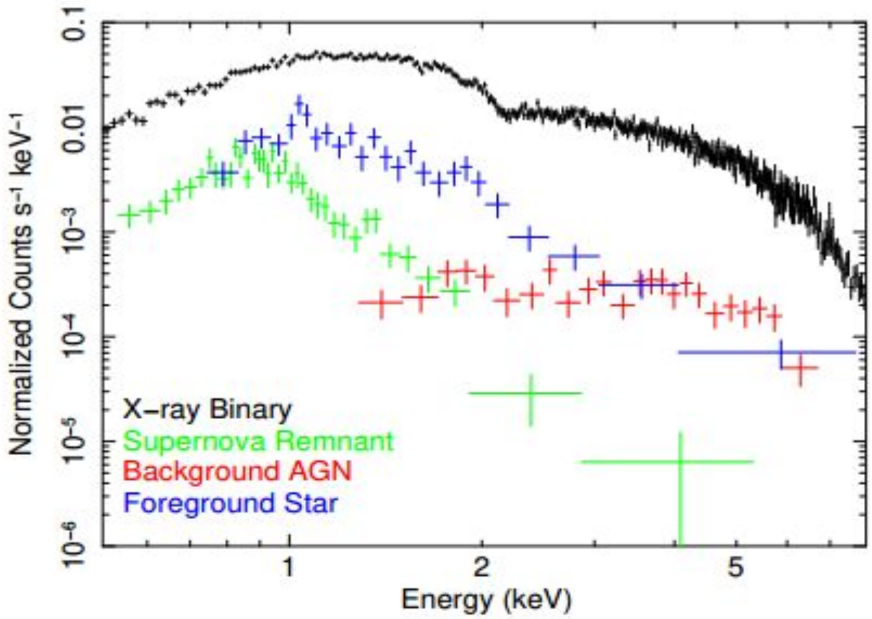


Figure 1. *Chandra* X-ray spectra of point source types detected in the direction of M31. The spectral shape of each source type is unique across the *Chandra* energy band of 0.5 – 8.0 keV, assuming sufficient source counts. In the low-count regime, advanced techniques such as ML are required to differentiate sources.

- **XRBs:** Generally well described by an absorbed power law with $\Gamma \sim 1.7$
- **AGNs:** Similar Γ to XRBs (unobscured AGNS), $\Gamma < 1..7$ (heavily absorbed AGNs)
- **SNRs:** Typically very soft sources . Shell-like & Crab-like: pulsar wind nebulae
- **fgStars:** X-ray emission due to flares from late type stars (e.g M-dwarfs)

Traditional classification of X-ray sources: Unique features to compact objects, colour-colour diagrams , observations in hard X-rays

In soft X-rays the distinguishing is difficult !

Machine learning for X-ray source classification

Solution for low energy-resolution X-ray data:

- Application of machine learning supervised algorithms to make optimal use of information in these energies

Supervised ML algorithms:

- Learn a relationship between a set of measurements and a target variable based on provided examples

Scientific goals of this work

- Development of an improved automated method for the distinguishing of extragalactic X-ray binaries based only on their X-ray emission
  - Improving the computation of XLF by avoiding contamination of non-XRB sources
  - Identifying new XRB candidates for follow up

X-Ray Data & Features

Sample dataset: "Catalogue of Chandra X-ray sources in M31" [Vulic et al.,2016]
Energy range: 0.5-8 keV   Total area: ~ 0.6 deg ^{2}

Classified sources: 163   [77 XRBs , 43 AGNs , 29 fgStars , 14 SNRs]
Unclassified sources: 780

Features

- 15 photon flux ratios
- Total photon flux 0.5 - 8 keV
- Mean observed energy
- Mean incident energy

Using of fluxes ratios for distance-independent features

Table 1. Summary of dataset properties

Feature Name	# classified	# unclassified
0.5 – 8.0 keV photon flux	163	780
0.5 – 2.0 keV photon flux fraction	163	749
2.0 – 8.0 keV photon flux fraction	153	744
0.5 – 1.7 keV photon flux fraction	163	736
1.7 – 2.8 keV photon flux fraction	152	679
2.8 – 8.0 keV photon flux fraction	147	723
0.5 – 1.5 keV photon flux fraction	162	728
1.5 – 2.5 keV photon flux fraction	156	684
2.5 – 8.0 keV photon flux fraction	149	731
0.5 – 1.0 keV photon flux fraction	155	634
1.0 – 2.0 keV photon flux fraction	163	719
2.0 – 4.0 keV photon flux fraction	148	686
4.0 – 6.0 keV photon flux fraction	139	636
6.0 – 8.0 keV photon flux fraction	121	513
0.5 – 7.0 keV photon flux fraction	163	779
2.0 – 7.0 keV photon flux fraction	152	742
Mean Observed Energy	163	768
Mean Incident Energy	156	664

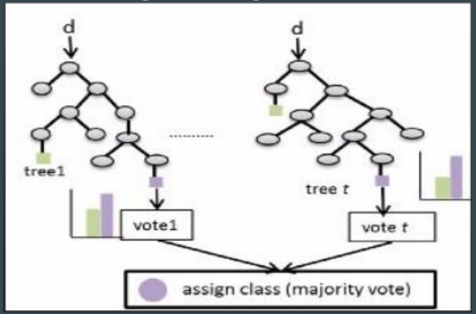
The number of classified and unclassified objects per feature varies because some objects have feature values set to zero due to a negative flux or energy being inferred from ACIS EXTRACT.

Algorithms

- **Logistic regression:** Assumes that classes are linearly separable in the features space and try to fit to probability of class membership - Similar to linear regression
- **Gaussian naive Bayes:** Assumes that all features are conditionally independent given the class label - Produces conditional class probabilities using Bayesian formulation
- **SVC:** Fits a separating hyperplane in the future space - Classification of features examples
- **Multi-layer perceptron:** A class of Neural Network - Possesses hidden layers that learn between the feature inputs and the fitted output - Learns well non - linear functions
- **Random Forest classifier:** A collection of a large number of decision trees
  - During the training process uses randomly-selected data subsets of the initial sample
  - Random subsets of features are used in each node of the decision tree

[Belgiu & Dragut , 2016]

Each tree in the forest suggest a class → Majority vote → Final prediction



Methodology

Classification scheme

Multiclass classification

- Classes: XRB, AGN, fgStar, SNR

Evaluation of the viability of classification across multiple object types

Binary classification

- Classes: XRB, non XRB

Primary goal: Identification of new XRBs candidates
2 classes: improvement of the algorithms performance (init. sample $<$ 200)

Algorithm implementation & evaluation

- Split classified samples 70% train - 30% test
- Basic optimization of the hyper-parameters
- ➢ K-fold cross-validation on the entire dataset
    - Dataset partitioned to k subsamples
    - k - 1 for training and k for test
    - cv score: Average accuracy

METRICS

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

RESULTS

Multiclass classification

Table 2. Algorithm Evaluation, multiclass case

Algorithm	Accuracy	Precision	Recall	CV Score
Logistic Regression	0.55	0.53	0.55	0.54 ± 0.04
Naive Bayes	0.57	0.54	0.57	0.52 ± 0.07
Support Vector Class.	0.49	0.43	0.49	0.55 ± 0.04
Random Forest (sklearn)	0.57	0.57	0.57	0.65 ± 0.06
Multi-layer Perceptron NN	0.57	0.60	0.57	0.52 ± 0.08
Random Forest (R)	0.61	0.61	0.60	0.66 ± 0.07

Table 3. Confusion matrix for sklearn random forest, multiclass case

		Actual Class				
		AGN	SNR	fgStar	XRB	Total
Predicted Class	AGN	5	0	2	2	9
	SNR	0	3	1	0	4
	fgStar	3	2	5	0	10
	XRB	6	1	4	15	26
	Total	14	6	12	17	49

- Generally poor performance metrics for the algorithms
- RF: Best performance
- MPNN : Poorest performance

- Most of misclassifications are from: fgStars & SNRs

WHY ?

- Underepresented classes
- Spectroscopically similar

RESULTS

Binary classification

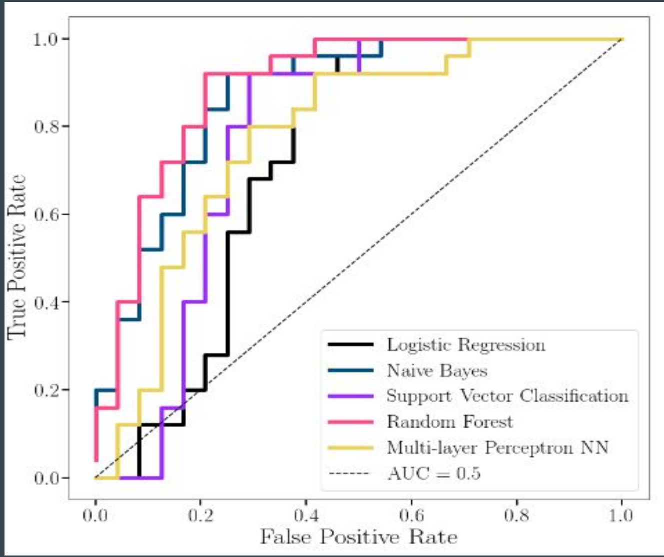
Table 5. Algorithm Evaluation, binary case

Algorithm	Accuracy	Precision	Recall	AUC*	CV Score
Logistic Regression	0.71	0.55	0.77	0.74	0.66 ± 0.06
Naive Bayes	0.73	0.58	0.77	0.85	0.74 ± 0.09
Support Vector Class.	0.71	0.55	0.77	0.85	0.71 ± 0.08
Random Forest (sklearn)	0.84	0.71	0.85	0.88	0.75 ± 0.05
Multi-layer Perceptron	0.61	0.46	0.63	0.62	0.72 ± 0.07
Random Forest (R)	0.86	0.75	0.86	0.89	0.79 ± 0.06

Table 6. Confusion matrix for sklearn random forest, binary case

		Actual Class XRB	Actual Class non-XRB	Total
Predicted Class	XRB	15	5	20
	non-XRB	2	27	29
	Total	17	32	49

- Accuracy is improved for all algorithms
- RF: Best performance with higher score than the multiclass approach
- For XRBs the number of misclassified objects is the same with multiclass approach
- The overall number of misclassifications is reduced



ROC curves

TPR= TP/TP+FN
FPR= FP/FP+TN
Ideal case : TPR =1 & FPR = 0

RF again has the best overall performance !

RESULTS

Classification validation by crossmatching

Goal: **Comparison the RF's classification strength with classifications based on other wavelengths. (e.g optical)**

1st step : Application of RF method to 780 X-ray sources (unseen data) [*Vulic et al. 2016*]

2nd step: Matched the 780 newly classified X-ray sources with those from 3 X-ray surveys in M31
41 matches in total

3rd step : Comparison of these 41 RF classified sources with the classifications of their optical counterparts in the PHAT survey

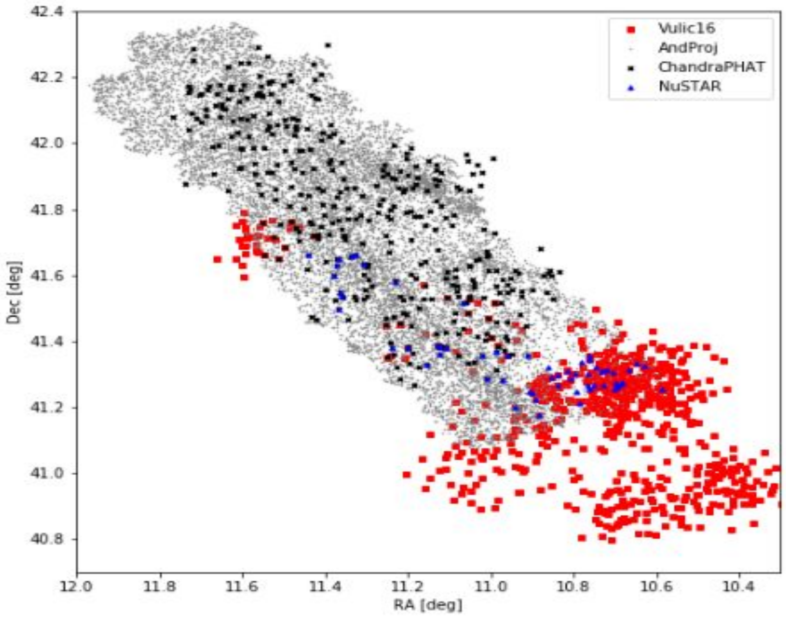


Figure 2. *Chandra Hubble* and *NuSTAR* sources in M31. Red squares: unclassified *Chandra* sources from Vulic et al. (2016), grey dots: Andromeda Project non-stellar (*HST*) sources from Johnson et al. (2015), black crosses: *Chandra*-PHAT sources from Williams et al. (2018), blue triangles: *NuSTAR-Chandra* sources from Lazzarini et al. (2018). Not shown here are sources in an

RESULTS

Compatibility criteria for X-ray and Optical Classification schemes		
X-ray source	**Compatible with optical source**	**Incompatible with optical source**
XRB	optical point sources, non-detection, star clusters, unknown	foreground stars,SNRs
Non XRB	All types of Hubble sources	Star clusters
AGN	optical point sources, non-detection,galaxies,unknown	Star clusters, foreground stars, SNRs
fgStar	foreground stars	All other types
SNR	SNRs	All other types

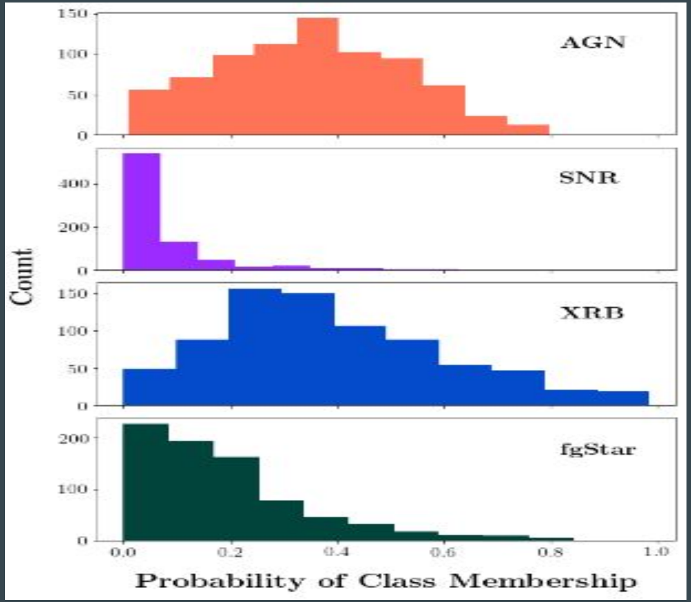
$$\text{Compatibility score} = \frac{\text{Numbers of objects with compatible classifications}}{\text{Total number of objects}}$$

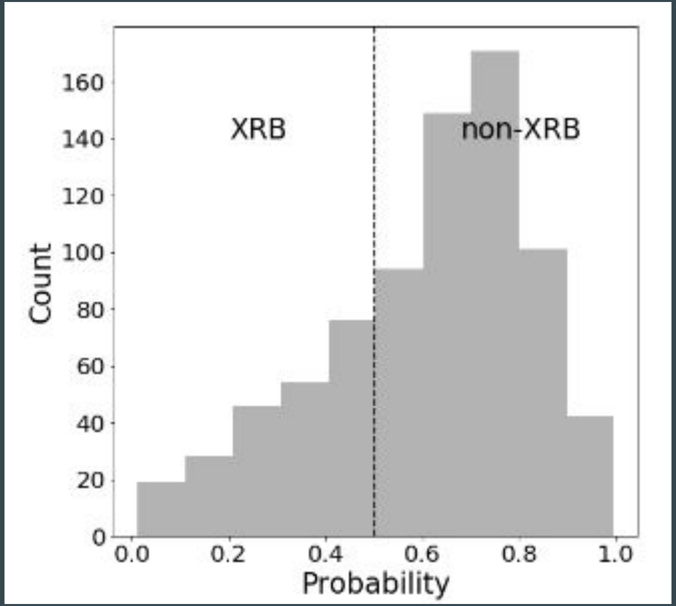
Compatibility score

31/41 ~ 91 %

RF classifications are in agreement with classifications based on non X-ray properties !

RESULTS





- SNR & fgStar : Peak at low Probability values
- XRB & AGN : Peak at higher values. Difficulty to separate these classes

19 XRBs candidates with P(XRB)> 90% !

P(XRB) in binary classification & multiclass classification

In very good agreement

16 XRBs candidates with P(XRB) > 90% !

RESULTS

Most important X-ray features

The most important features during the training of RF classifier are the Photon Flux ratios for the X-ray bands:

- 1.7-2.8 keV
- 0.5-1 keV
- 2.0-4.0 keV
- 2.0-7.0 keV

Not expected result !

Less common bands in traditional hardness ratio analyses !

Narrower bands are expected to be less useful

Detailed interpretation of these bands in a future work

TAKE HOME MESSAGE

- RF forest classifier is the best among other supervised algorithms , with an accuracy ~85 % (binary case)
- 16 new strong (P(XRB) > 90 %) XRB candidates are suitable for follow up
- Cross-matching previously unclassified sources X-ray sources with sources classified using PHAT resulted in compatibility score ~ 91 %
- The narrower and less commonly used bands as 1.7-2.8 , 0.5-1.0 , 2.0-4.0 & 2.0-7.0 keV photon flux ratios are the most important for the classification