Machine-learning Classification of X-ray Sources in the Era of Modern X-ray Observatories.

Oleg Kargaltsev[1], Hui Yang[1], Jeremy Hare[2], Steven Chen[1]
The George Washington University[1], NASA GSFC[2]

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

GitHub SCAN ME!
CSC Training Dataset

The MUltiWavelength CLASSification Pipeline (MUWCLASS)

Yang, Hare, Kargaltsev, et al. (2022)

Training Dataset (TD)

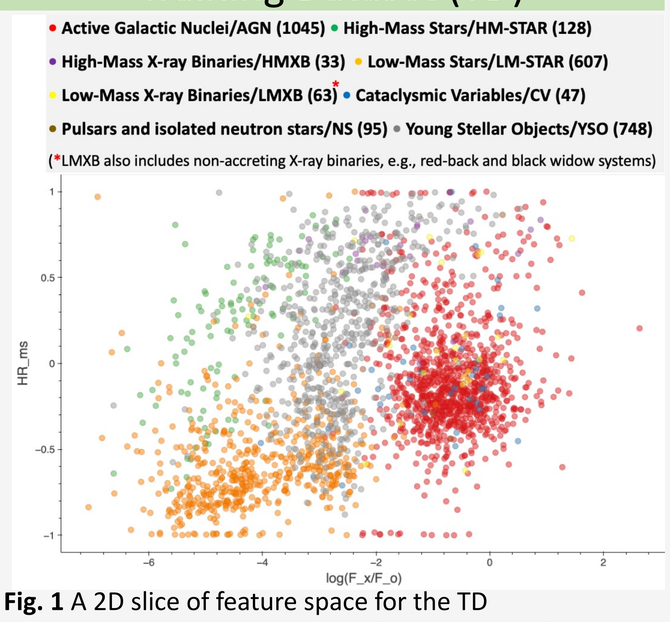


Fig. 1 A 2D slice of feature space for the TD

- Each source can have up to 29 MW features: ■ X-ray fluxes, hardness ratios, X-ray variability (CSCv2) ■ optical (Gaia) ■ NIR (2MASS) ■ MIR (WISE) photometry & colors. (See Fig. 3)
- Explore the TD yourself using the visualization GUI at https://home.gwu.edu/~kargaltsev/XCLASS/ (Yang+2021)

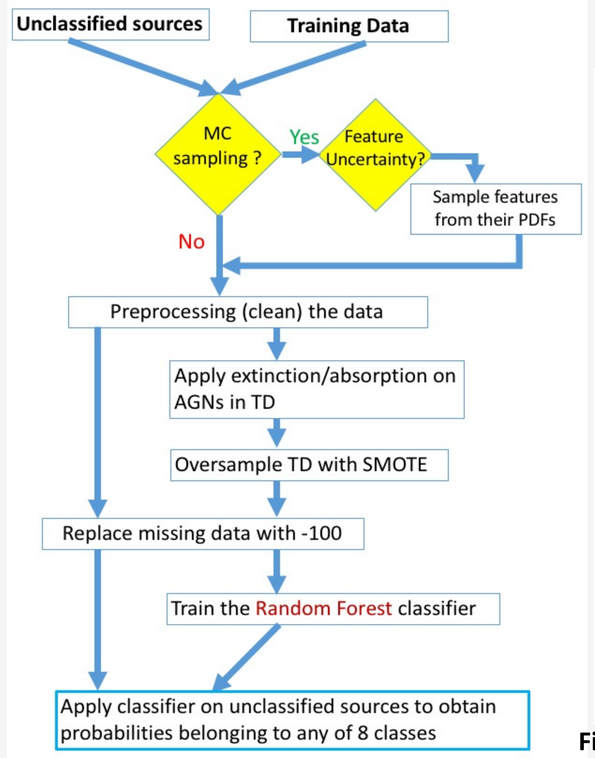


Fig. 2 The workflow chart of MUWCLASS pipeline

MUWCLASS performance

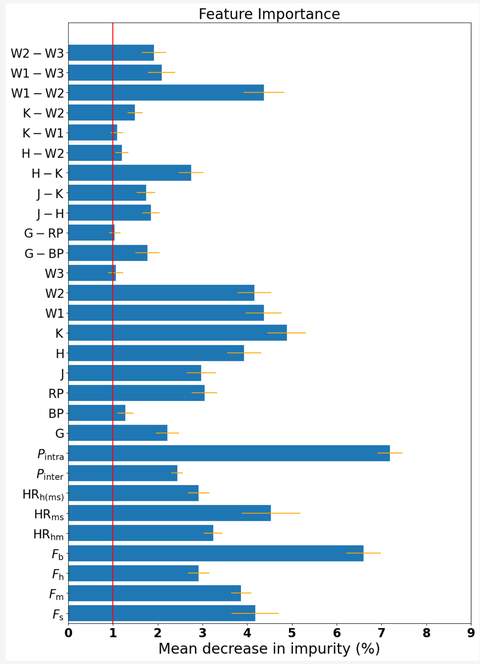


Fig. 3 The feature importance

- Feature importance = how often a particular feature is being used in the classification process.

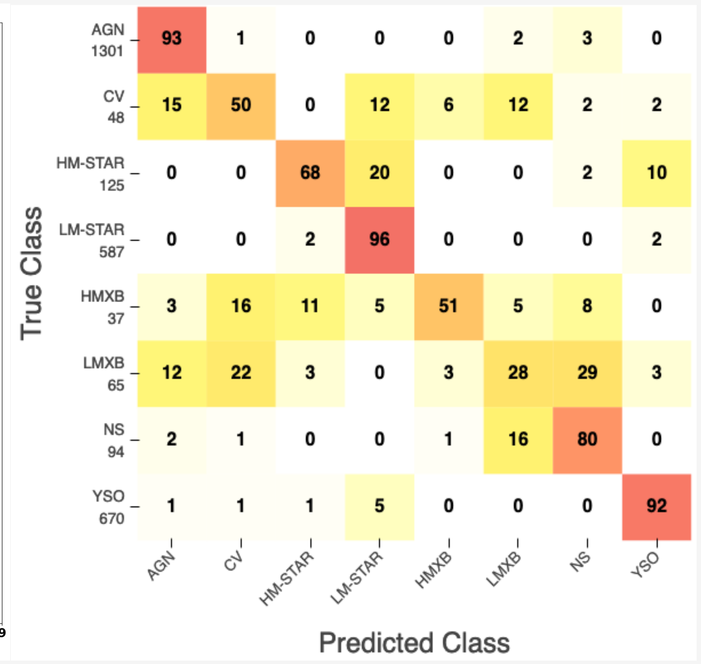


Fig. 4 Normalized confusion (**performance**) matrix

- A more diagonal matrix → better performance
- Not all classes are classified equally well.

Recent Developments

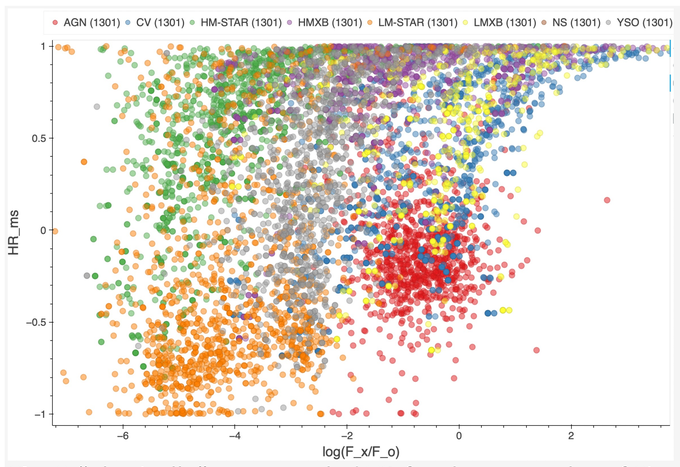


Fig. 5 "physically" oversampled TD for the same plot of Fig. 1.

- The TD is imbalanced (see the # of sources for each class in Fig. 1).
- We produce synthetic sources by sampling reddening parameters from those of TD and applying it on the less-populated (excluding AGN) class.
- This oversampling is more realistic/"physical" than other algorithms (e.g., SMOTE), and produce a fainter population of sources.

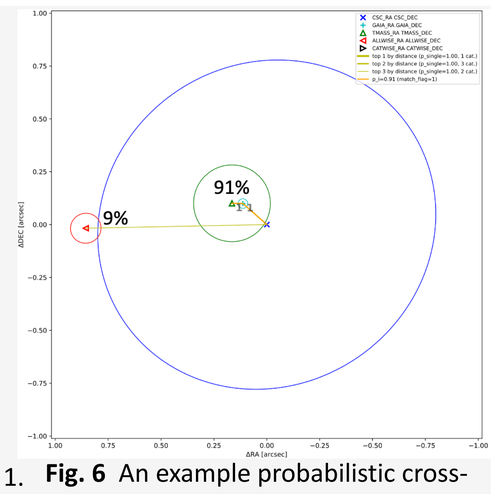


Fig. 6 An example probabilistic cross-matching of X-ray sources with multi-wavelength catalogs using NWAY algorithm (Salvato+2018).

Results: Classifications of CSC sources from Galactic Plane (Yang et al. 2022)

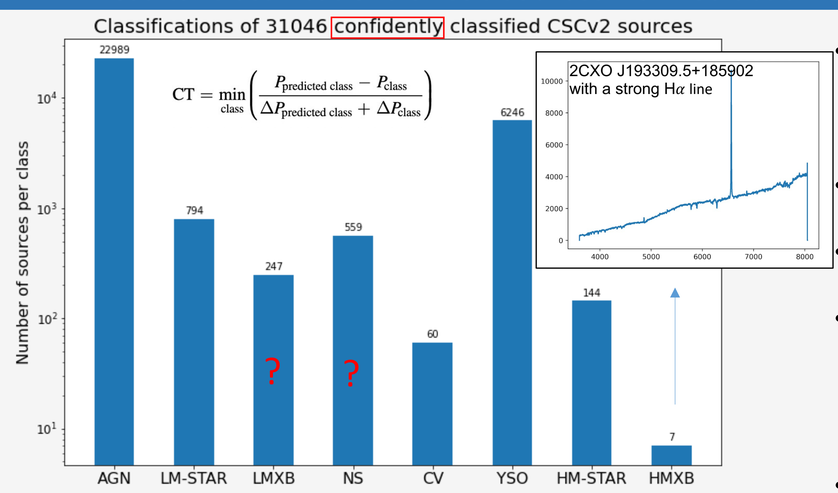


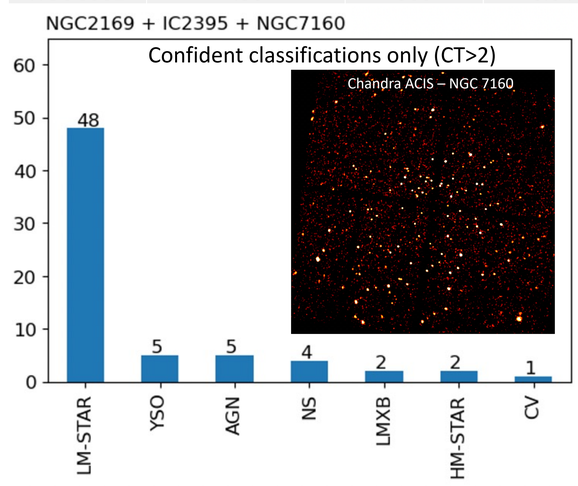
Fig. 7 The classification breakdown of confidently classified CSCv2 sources from "good" sample.

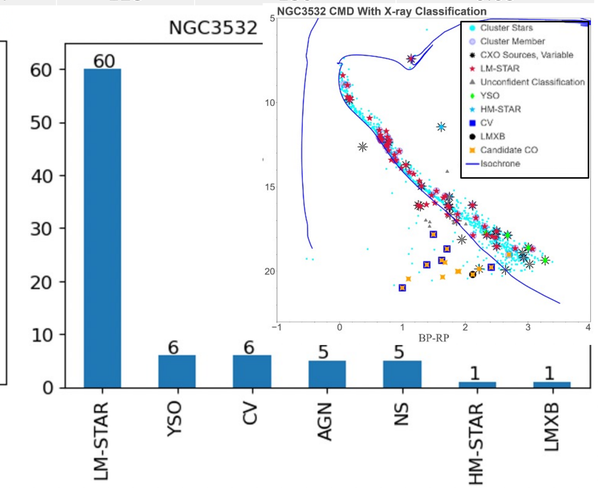
Classification confidences (CTs) are obtained by Monte-Carlo sampling of measurement uncertainties for the source features.

- "Good" Sample: ~66,400 sources (20%) with significant detections (S/N≥3) and accurate positions (PU <1").
- About 50% are confidently (CT>2) classified.
- Some HMXBs are already independently confirmed.
- For many LMXBs counterparts are too faint to be detected by MW surveys, hence LMXBs become confused with the NS class.
- Sources that are too faint to be detected by MW surveys (e.g., flaring M-dwarfs, absorbed AGN) can classified as NS/LMXBs.

Results: X-ray source in open clusters (Chen et al. 2023)

Cluster	Distance (pc)	Age (Myr)	E(B-V)	N_CXO	T_CXO (ks)	zeta (s/pc^2)
NGC 2169	959	7	0.23	15	63	0.12
IC 2395	717	9	0.16	40	64	0.12
NGC 7160	800	16-25	0.25	22	69	0.11
NGC 3532	450	300	0.04	128	130	0.63





Results: Classifications of X-ray (CXO) sources in the fields of unidentified 4FGL sources

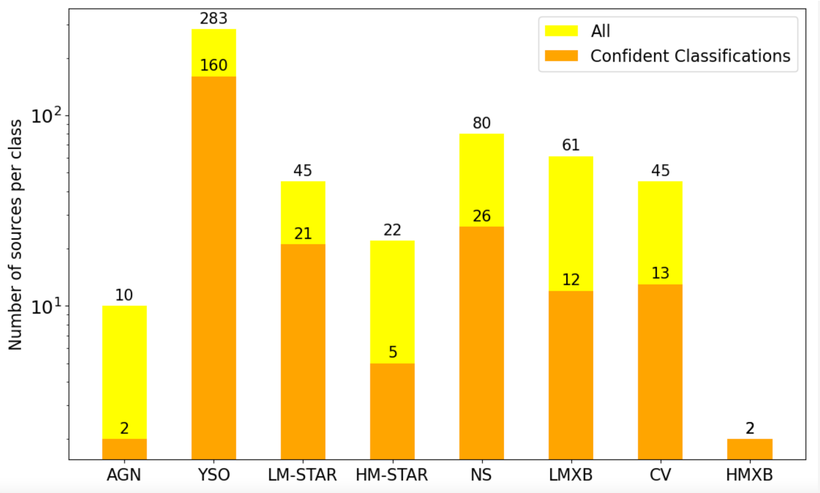
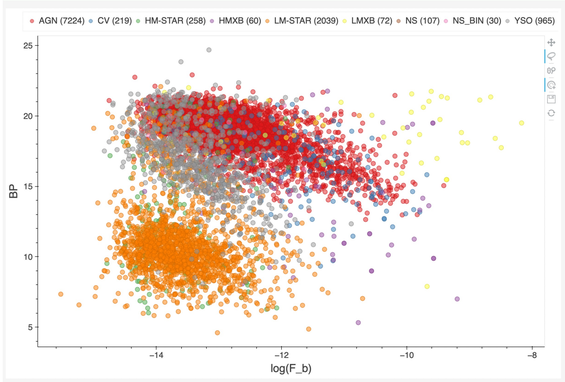
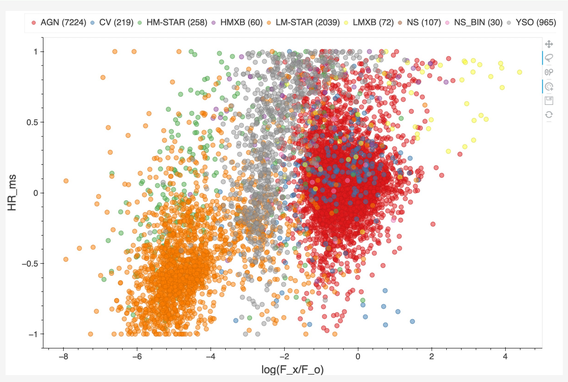


Fig. 8 The classification breakdown of CSCv2 sources within error ellipses of 37 unassociated 4FGL-DR3 sources

- We classify 26 NS, 2 AGNs, and 2 HMXB as γ-ray emitter candidates associated with 37 unassociated 4FGL-DR3 sources.
- We also classify 160 YSO most of them near SFR or young clusters.
- Higher IR background in the plane leads to a bias in the TD were the AGN are taken from the surveys outside the Galactic plane. This can cause AGNs to be incorrectly classified as NSs.
- Deeper NIR and radio observations of the NS candidates are needed to verify the classifications.
- Yang et al. in prep.

Ongoing work: XMM-Newton training data set

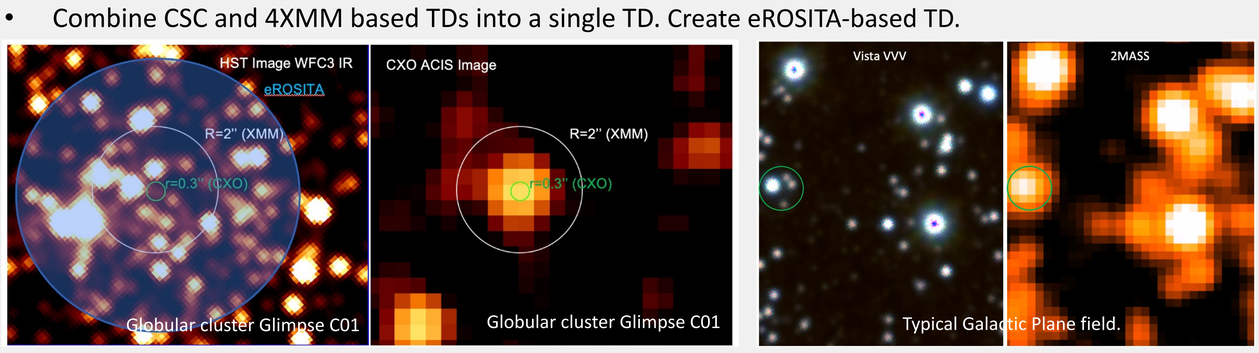




- Due to the less certain source positions, verifications of counterparts to 4XMM sources with firm literature-based classifications takes more efforts compared to CSC. Currently, 4XMM TD has ~10,000 sources.
- We will be using this TD to classify 4XMM sources in the Galactic Plane and further explore interesting classifications within 4FGL sources, SNRs, etc.
- We plan to add radio data once MeerKat GPS catalog(s) become available.

Outlook:

- Create a "living" database of X-ray sources of firmly known types to be used as an ultimate training dataset which is constantly expanding and driven by the community efforts.
- Create a database of confidently classified (using ML) sources suitable to population studies.
- Include additional features: optical variability, radio flux, distances, luminosities, etc.
- Replace all-sky MW surveys with more sensitive surveys covering substantial parts of the Galactic plane (Vista VVV, Glimpse, UKIDSS, VPHAS+, etc).
- Develop more reliable probabilistic source matching and incorporate the MW matching probabilities into the classification probabilities.
- Combine CSC and 4XMM based TDs into a single TD. Create eROSITA-based TD.



References: • Yang, H., Hare, J., Kargaltsev, O., et al. 2022, ApJ, 941, 104. • Yang, H., et al. 2021, RNAAS, 5, 102 • Salvato, M., et al. 2018, MNARS, 473, 4937 • Chen, S., Kargaltsev, O., Yang, H., et al. 2023, ApJ, 948, 31.
Acknowledgement: This research is partly supported by NASA through the Chandra X-ray Observatory award *AR3-14017X, AR9-20005, and AR8-19008X,* AR0-21007X and the Astrophysics Data Analysis Program award 80NSSC19K0576.