Draft version May 6, 2026
Typeset using LaTeX **twocolumn** style in AASTeX7.0.1

XClass: An Automated Multiwavelength Machine-Learning Pipeline for Classification of Extragalactic X-ray Sources. I. Pipeline Description

Blagoy Rangelov [1], Nicholas Moore,[1] Haley Dutton,[1] and Erika Marentes [1]

[1] *Department of Physics, Texas State University, 601 University Drive, San Marcos, TX 78666, USA*

ABSTRACT

Most X-ray sources detected in nearby galaxies by the *Chandra X-ray Observatory* lack astrophysical classifications. We present XClass (X-ray Classifier for extragalactic sources), an end-to-end machine-learning pipeline that classifies extragalactic X-ray point sources into seven classes: AGN, LMXBs, HMXBs, CVs, low-mass and high-mass foreground stars, and supernova remnants. The key challenge is photometric heterogeneity: training sources are predominantly Galactic with photometry from wide-field surveys (PanSTARRS, *Gaia*, 2MASS), while extragalactic targets require *Hubble Space Telescope* (HST) imaging in a disjoint filter system. We address this through a spectral energy distribution (SED) translation that fits class-appropriate spectral models to each training source and convolves the best-fit model through HST filter curves, producing synthetic magnitudes in a common feature space. The classifier uses an asymmetric two-stage Random Forest: Stage 1 separates broad categories (AGN, X-ray binaries, SNRs, stars) and Stage 2 resolves X-ray binaries into LMXBs and HMXBs using an augmented feature vector that includes Stage 1 probabilities. The training set is assembled from ten Galactic catalogs and extragalactic SNR catalogs, cross-matched with the Chandra Source Catalog v2.0. Features include X-ray hardness ratios, SED-translated HST colors, X-ray-to-optical flux ratios, and *Gaia* astrometric properties. We restrict the training set to sources with at least one optical magnitude, avoiding imputation artifacts; the pipeline achieves 99.6% accuracy and balanced accuracy of 0.91 on the resulting 11,374-source optical baseline, with excellent calibration (ECE = 0.002). XClass is modular, generalizable to any HST filter configuration, and will be applied to M31 and M33 in a companion paper.

Keywords: X-ray sources (1822) — Classification (1907) — Random Forests (1935) — X-ray binary stars (1811) — Active galactic nuclei (16) — Catalogs (205) — Supernova remnants (1667) — Cataclysmic variable stars (203) — X-ray surveys (1824) — Astrostatistics tools (1887)

1. INTRODUCTION

A census of extragalactic X-ray source populations is essential for addressing questions ranging from X-ray binary evolution to the star formation histories of host galaxies (e.g., G. Fabbiano 2006; M. Gilfanov 2004). The *Chandra X-ray Observatory* has enabled the detection of thousands of discrete X-ray sources in galaxies within ∼10 Mpc (e.g., I. N. Evans et al. 2010, 2020), yet the vast majority remain unclassified. In M31 and M33 alone, ∼3,000 X-ray sources have been cataloged, but fewer than ∼1/4 have reliable astrophysical classifications (e.g., H. Stiele et al. 2011; B. F. Williams et al. 2018; R. Tüllmann et al. 2011).

In the absence of spectra, classification must rely on broadband photometric and X-ray properties — hardness ratios, colors, and X-ray–to–optical flux ratios (e.g., D. L. Kaplan et al. 2006; Z. Misanovic et al. 2010; N. J. Brassington et al. 2012). However, traditional two-parameter diagnostic approaches become impractical for thousands of sources in a high-dimensional feature space and do not provide quantified classification confidences. Machine-learning (ML) methods such as Random Forests (L. Breiman 2001) offer a solution, since they operate on many features simultaneously and provide probabilistic class assignments. Several groups have applied ML to X-ray source classification, primarily in Galactic fields (K. K. Lo et al. 2014; H. Tranin et al. 2022). H. Yang et al. (2022) presented MUWCLASS, a Random Forest pipeline using X-ray, optical, NIR, and IR photometry from the Chandra Source Catalog (CSC), PanSTARRS, *Gaia*, 2MASS, and *WISE*, which

was used to classify ∼66,000 CSC v2.0 sources and subsequently applied to NGC 3532 (S. Chen et al. 2023) and to 13 unidentified *Fermi* LAT sources (B. Rangelov et al. 2024). In the extragalactic domain, R. M. Arnason et al. (2020) used Random Forests with X-ray properties alone to perform binary XRB/non-XRB classification of 943 *Chandra* sources in M31; however, that work did not incorporate optical or NIR photometry.

All of the above efforts share a common limitation: the training and application data use the same photometric system. Extending ML classification to extragalactic sources using multiwavelength (MW) data introduces a fundamental challenge — photometric heterogeneity between the training and application domains. Confidently classified training sources are overwhelmingly Galactic, with counterparts characterized by wide-field surveys (PanSTARRS, *Gaia*, 2MASS), while extragalactic targets require *Hubble Space Telescope* (HST) imaging to resolve faint counterparts in crowded fields. Given that the HST filter system is entirely disjoint from the ground-based sets, direct feature-space comparison is not possible.

In this paper, we present **XClass** (**X**-ray **Class**ifier for extragalactic sources), an end-to-end ML pipeline for classifying extragalactic X-ray point sources into seven classes: active galactic nuclei (AGN), low mass X-ray binaries (LMXBs), high mass X-ray binaries (HMXBs), cataclsysmic variables (CVs), low-mass foreground stars (LM-STARs), high-mass foreground stars (HM-STARs), and supernova remnants (SNRs). XClass builds on the conceptual framework of MUWCLASS but is redesigned for the extragalactic context. The key innovation is a *spectral energy distribution (SED) translation* procedure: for each training source, a class-appropriate physical SED model is fit to the ground-based photometry and convolved through HST filter curves, producing synthetic HST magnitudes that place training and application data into a common feature space. The approach is general, accommodating any combination of input surveys and output HST filters.

In addition to the SED translation, XClass incorporates: (1) an asymmetric two-stage Random Forest that separates broad categories before resolving the LMXB/HMXB distinction; (2) a training dataset assembled from ten literature catalogs of Galactic sources, augmented with extragalactic SNR catalogs; (3) cross-matching with CSC 2.0 and photometric queries to PanSTARRS DR2, *Gaia* DR3, and 2MASS; and (4) counterpart matching with the Hubble Source Catalog v3 for target galaxies. To our knowledge, XClass is among the first ML pipelines designed specifically for extragalactic X-ray source classification using MW data across heterogeneous filter systems. Given the growing *Chandra* and HST archives — and the advent of *Athena* (K. Nandra et al. 2013), AXIS (R. Mushotzky 2018), Lynx (J. A. Gaskin et al. 2019), the Einstein Probe (W. Yuan et al. 2022), the *Vera C. Rubin Observatory* (Ž. Ivezić et al. 2019), and the *Roman Space Telescope* — such a tool is both timely and necessary.

The rest of the paper is organized as follows. Section 2 describes the training dataset. Section 3 presents the SED translation procedure. Section 4 covers feature engineering and the two-stage classifier. Section 5 evaluates classification performance. We discuss design decisions and limitations in Section 6 and summarize in Section 7.

2. TRAINING DATA

The XClass training dataset is assembled from literature catalogs of confidently classified X-ray sources, cross-matched with detections in the Chandra Source Catalog, and augmented with ground-based MW photometry.

2.1. *Source Classes and Label Catalogs*

XClass classifies sources into seven astrophysical classes, selected to represent the dominant X-ray point-source populations in nearby spiral galaxies. The class scheme builds on the Galactic MUWCLASS pipeline (H. Yang et al. 2022), omitting isolated neutron stars (too faint at extragalactic distances) and young stellar objects (unresolved in external galaxies), and absorbing Wolf-Rayet stars into the high-mass star class. Training labels were drawn from ten literature catalogs queried from VizieR (F. Ochsenbein et al. 2000) via `astroquery` (A. Ginsburg et al. 2019), plus three extragalactic SNR catalogs.

AGN — background sources seen in projection or nuclear sources within the target galaxy — were drawn from the Véron-Cetty & Véron catalog of quasars and active nuclei (M.-P. Véron-Cetty & P. Véron 2010). AGN are the most numerous class in both the training set and typical extragalactic X-ray catalogs (H. Stiele et al. 2011; B. F. Williams et al. 2018).

LMXBs, containing a compact object accreting from a low-mass ($\lesssim 1\,M_\odot$) donor via Roche lobe overflow, trace the old stellar population (M. Gilfanov 2004). Labels were compiled from the Milky Way LMXB catalog of Q. Z. Liu et al. (2007) and the X-ray binary tables of H. Ritter & U. Kolb (2003).

HMXBs, with compact objects accreting from massive ($\gtrsim 8\,M_\odot$) OB or Be-star donors, are associated with recent star formation (H.-J. Grimm et al. 2003). Labels were taken from the HMXB catalog of Q. Z. Liu et al. (2006).

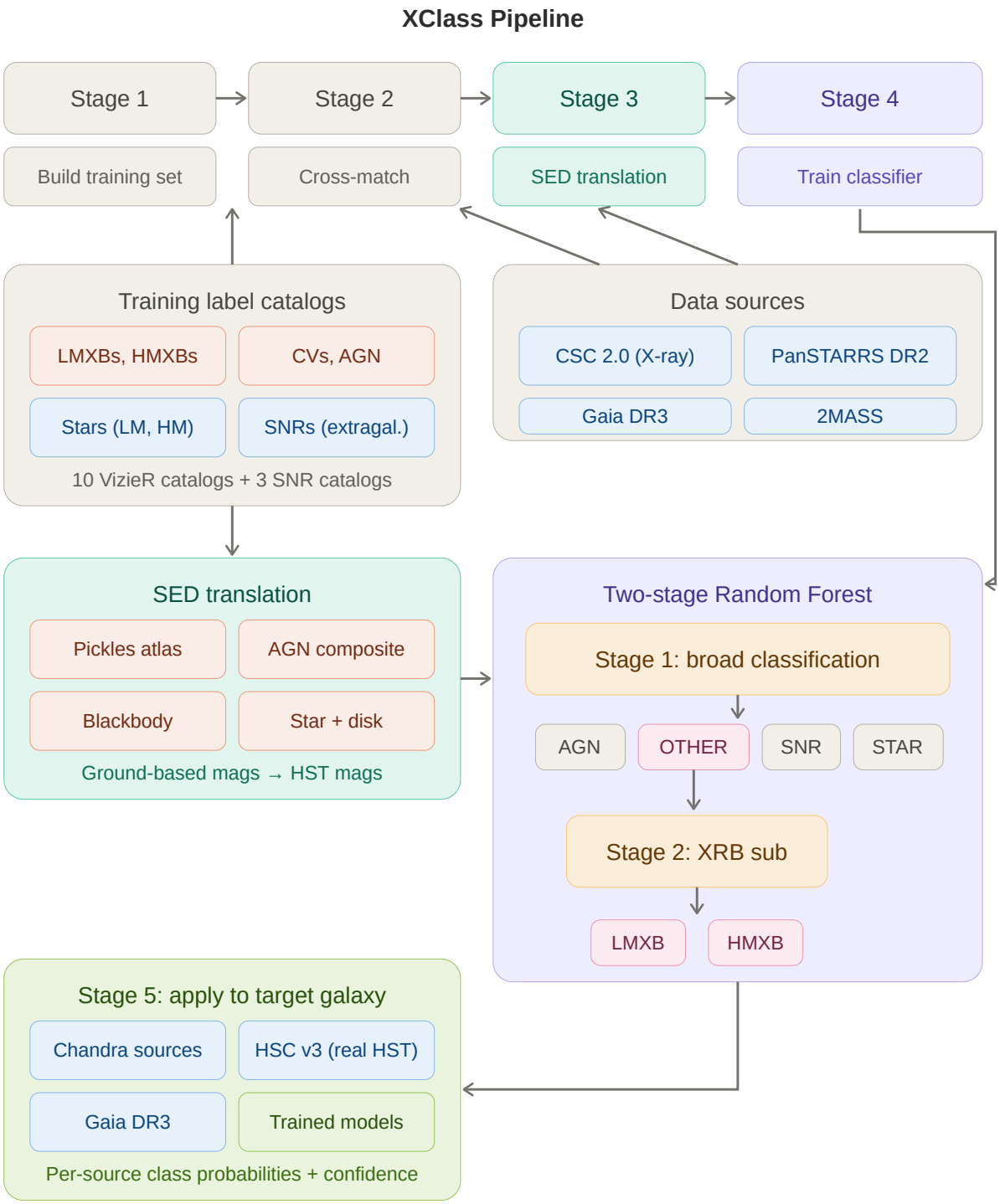

Figure 1. End-to-end schematic of the XClass pipeline. The five stages are shown from left to right: (1) training-set construction from literature catalogs, (2) cross-matching with the Chandra Source Catalog and ground-based photometric surveys, (3) SED translation from ground-based to HST photometric features (Section 3), (4) two-stage Random Forest training (Section 4), and (5) application to target galaxies using real HST photometry from the Hubble Source Catalog. The two-stage classifier architecture is shown as an inset: Stage 1 classifies sources into four broad groups (AGN, OTHER, SNR, STAR), and Stage 2 resolves OTHER into LMXB or HMXB.

CVs are white-dwarf binaries accreting from low-mass companions. Although intrinsically faint in X-rays, they appear as foreground contaminants. Labels were drawn from the CV tables of H. Ritter & U. Kolb (2003) and the atlas of R. A. Downes et al. (2001).

LM-STARs of spectral types A–M emit X-rays through coronal activity and are ubiquitous foreground contaminants. **HM-STARs** of spectral types O–B (including Wolf-Rayet stars) produce X-rays via stellar winds. Stellar labels were obtained primarily from the MK spectral classification catalog of B. A. Skiff (2014), with O/B types assigned to HM-STAR and later types to LM-STAR; ambiguous types were excluded. These were supplemented by APOGEE-2 DR16 (S. R. Majewski et al. 2017) sources (with valid $T_{\rm eff}$, $\log g$, and radial-velocity scatter ≤ 1.0 km s^{-1}) as LM-STARs, and by Wolf-Rayet stars from K. A. van der Hucht (2001, 2006) as HM-STARs.

SNRs are the sole class trained on extragalactic data, since Galactic SNRs observed by *Chandra* are typically highly extended and poorly representative of the compact remnants seen in external galaxies. We use SNR catalogs from M33 (K. S. Long et al. 2010), NGC 6946 (K. S. Long et al. 2019), and M83 (W. P. Blair et al. 2012), cross-matched to *Chandra* detections at a fixed radius of 3.″0. SNRs are excluded from the SED translation step (Section 3) because their training features rely only on HST photometry when available, or on X-ray properties alone.

2.2. *Deduplication*

After concatenating all catalog entries, deduplication was applied within each source class using a Union-Find data structure with two merging criteria: (1) rows with identical normalized names were merged unconditionally, and (2) positional pairs within 5″ were merged if they share a name, either name is empty, or their separation is $\leq 1''$. For each merged group, the canonical position is the median R.A. and decl., and the canonical name is the shortest non-empty string. SNR deduplication uses a tighter 2″ radius with priority given to X-ray detections over optical or radio.

2.3. *Cross-matching with the Chandra Source Catalog*

The deduplicated training-set positions were cross-matched against CSC 2.0 (I. N. Evans et al. 2010, 2020), which contains ∼317,000 unique X-ray sources with positions, fluxes in four energy bands (soft: 0.5–1.2 keV, medium: 1.2–2.0 keV, hard: 2.0–7.0 keV, broad: 0.5–7.0 keV), and positional uncertainties. The effective positional uncertainty for each Chandra source was computed from the error ellipse semi-axes r_0 and r_1 as $\sigma_{\rm eff} = \sqrt{(r_0^2 + r_1^2)/2}$. The search radius was set to $R_{\rm search} = \max(3\,\sigma_{\rm eff},\ 0.''5)$, and only sources with detection significance $\geq 3\sigma$ were included.

For each matched pair, we compute the normalized separation $\nu = \theta/\sigma_{\rm eff}$, where θ is the angular separation. Matches with $\nu \leq 2.0$ are flagged as "secure," those with $2.0 < \nu \leq 3.0$ as "ambiguous," and those with $\nu > 3.0$ are excluded. The normalized separation is retained and can be used as a classification feature (Section 4). The X-ray flux uncertainty for each band is estimated as $\sigma_{F_x} = (F_{x,\rm hi} - F_{x,\rm lo})/2$.

2.4. *Ground-Based Photometry*

For each training source matched to a Chandra detection, MW photometry was queried from three surveys: PanSTARRS DR2 (K. C. Chambers et al. 2016) (*grizy*; Dec $> -30°$), *Gaia* DR3 (Gaia Collaboration et al. 2023) (G, BP, RP, plus parallax and proper motion), and 2MASS (M. F. Skrutskie et al. 2006) (JHK_s). The nearest counterpart within 5″ was selected for each source.

2.5. *Training Set Summary*

Table 1 summarizes the number of sources per class at each stage. We note that the training set exhibits substantial class imbalance, with AGN dominating and compact-object classes (LMXB, HMXB, CV) being underrepresented. This motivates the use of balanced class weighting during classifier training (Section 4).

3. SED TRANSLATION

The central methodological innovation of XClass is the SED translation step, which bridges the photometric gap between the ground-based surveys used to characterize training sources and the HST photometry available for extragalactic targets. Training sources have photometry from PanSTARRS (*grizy*), *Gaia* (G, BP, RP), and 2MASS (JHK_s), while extragalactic targets are observed through entirely different HST filters (e.g., F275W, F336W, F475W, F814W, F110W, F160W for the PHAT survey; J. J. Dalcanton et al. 2012). Simple linear color transformations (e.g., M. Sirianni et al. 2005) are inadequate because they are source-type dependent — the correction between PanSTARRS i and HST F814W differs for a power-law AGN, a K-type star, and a flat accretion disk — thus introducing systematic biases for precisely the classes the classifier must distinguish. The SED translation resolves this by fitting a physically motivated spectral model to each training source and convolving the best-fit model through the HST filter curves, producing synthetic HST magnitudes that preserve class-discriminating color information.

Table 1. Training dataset summary by class. Columns show the number of sources at each pipeline stage: (1) after catalog compilation and deduplication, (2) after cross-matching with CSC 2.0 ($\geq 3\sigma$ detections), (3) after SED translation (sources with ≥ 3 photometric bands), and (4) final count entering the classifier, restricted to sources with ≥ 1 non-NaN predicted HST magnitude (see Section 4.2).

Class	Catalogs[a]	CSC Match	SED ($\geq$3)	Classifier[b]
AGN	168,314	5,954	5,231	5,231
LMXB	255	100	82	86
HMXB	114	37	36	39
CV	1,429	98	87	86
LM-STAR	947,698	4,693	4,615	4,610
HM-STAR	58,085	1,232	1,215	1,214
SNR	639	···	···	108[c]
Total	1,176,534	12,114	11,266	11,374

[a] Counts before any X-ray cross-matching; the large stellar totals reflect the full Skiff and APOGEE catalogs prior to positional matching with Chandra.

[b] Sources with ≥ 1 non-NaN predicted HST magnitude from SED translation or, for SNRs, from direct HSC photometry. Sources with zero optical features are excluded to avoid median-imputation artifacts (Section 4.2).

[c] SNRs bypass SED translation; the 108 sources reflect those with ≥ 1 PHAT-filter detection in the Hubble Source Catalog.

3.1. *Physical SED Models and Class Assignment*

Filter transmission curves for both input (ground-based) and output (HST) systems were obtained from the SVO Filter Profile Service (C. Rodrigo & E. Solano 2020). The HST set covers ∼40 filters across ACS/WFC, WFC3/UVIS, and WFC3/IR. For a specific target galaxy, only the relevant subset can be used as classification features. All SED models are evaluated on a common wavelength grid spanning 100–25,000 Å in 5 Å steps. Four model families are employed.

Pickles stellar atlas. Empirical spectra from the A. J. Pickles (1998) UV–optical–IR library, covering 27 spectral types from O5 to M6 (loaded from the STScI CDBS archive). For sources with a known spectral type, the nearest library type is selected; otherwise, all 27 templates are searched.

AGN composite. The mean SDSS quasar spectrum of D. E. Vanden Berk et al. (2001), covering the UV through NIR with the characteristic power-law continuum and broad emission lines.

Power-law. $f_\nu(\lambda) \propto (\lambda_{\rm ref}/\lambda)^\alpha$ with $\lambda_{\rm ref} = 5500$ Å and $\alpha \in [-2.5, +1.5]$ in steps of 0.05 (81 grid points). This is used as a fallback for AGN.

Blackbody. A Planck function with a temperature grid spanning 2,000–9,750 K in 250 K steps and 10,000–50,000 K in 1,000 K steps (73 grid points). This is used as a fallback for stars and X-ray binaries.

Two-component star + disk. For X-ray binaries and CVs, the optical emission is modeled as

$$f_\nu(\lambda) = (1 - f_{\rm disk}) \cdot f_{\nu,{\rm star}}(\lambda) + f_{\rm disk} \cdot f_{\nu,{\rm disk}}(\lambda), \quad (1)$$

where $f_{\nu,{\rm star}}$ is a Pickles template (K5 for LMXBs/CVs, B2 for HMXBs; P. A. Charles & M. J. Coe 2006) and $f_{\nu,{\rm disk}}$ is a flat (f_ν = const) spectrum representing a steady-state accretion disk. The disk fraction is varied over $f_{\rm disk} \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

Each class is assigned a primary and fallback model as summarized in Table 2. SNRs are excluded from SED translation entirely (see Section 2.1).

3.2. *Fitting Procedure*

The SED fit is applied independently to each training source with at least three valid photometric bands.

Step 1: Photometric conversion. Magnitudes were converted to flux densities f_ν (erg s^{-1} cm^{-2} Hz^{-1}) using the AB zero point (3631 Jy) for PanSTARRS and

Table 2. SED model assignment by source class. The primary model is used first; if it fails, the fallback model is attempted. SNRs bypass SED translation and use direct HST photometry when available.

Class	Primary Model	Fallback Model
AGN	AGN composite	Power-law
LMXB	Two-component (K5 + disk)	Blackbody
HMXB	Two-component (B2 + disk)	Blackbody
CV	Two-component (K5 + disk)	Blackbody
LM-STAR	Pickles atlas	Blackbody
HM-STAR	Pickles atlas	Blackbody

Gaia, and Vega zero points for 2MASS (Z_ν = 1594, 1024, 666.7 Jy for J, H, K_s). Uncertainties were propagated as $\sigma_{f_\nu} = f_\nu \cdot (\ln 10/2.5) \cdot \sigma_m$; when unavailable, 10% of the flux was adopted.

Step 2: Normalization. The band with the highest observed f_ν was selected as the normalization anchor.

Step 3: Grid search. For each model parameter (temperature, spectral index, disk fraction, or spectral type), the model was normalized to match the anchor-band flux and the reduced chi-squared was evaluated as

$$\chi^2_{\rm red} = \frac{1}{N_{\rm bands} - 1} \sum_{k \neq {\rm norm}} \left(\frac{f^{\rm obs}_{\nu,k} - \langle f_\nu \rangle^{\rm model}_k}{\sigma_{f_{\nu,k}}} \right)^2, \quad (2)$$

where the synthetic flux through filter k is calculated as

$$\langle f_\nu \rangle_k = \frac{\int f_\nu(\lambda)\, T_k(\lambda)\, d\lambda}{\int T_k(\lambda)\, d\lambda}. \quad (3)$$

Step 4: Best model selection. The grid point minimizing $\chi^2_{\rm red}$ was selected. If the primary model fails, the fallback is attempted; sources for which both fail are excluded.

Step 5: HST magnitude prediction. The best-fit SED was convolved through each target HST filter via Equation (3), yielding predicted AB magnitudes $m_{\rm pred} = -2.5 \log_{10}(\langle f_\nu \rangle) - 48.60$. These translated magnitudes place training sources into the same feature space as the HST-observed extragalactic application data.

3.3. *Uncertainty Propagation*

The uncertainty on each predicted magnitude is estimated as

$$\sigma_{\rm base} = \min\left(0.1 \cdot \sqrt{\chi^2_{\rm red}},\ 1.0 \text{ mag}\right). \quad (4)$$

For UV filters (F275W, F336W), an additional 0.4 mag systematic is added in quadrature:

$$\sigma_{\rm UV} = \sqrt{\sigma^2_{\rm base} + (0.4 \text{ mag})^2}. \quad (5)$$

This reflects the absence of ground-based constraints shortward of ∼4800 Å (PanSTARRS g-band); thus, UV predictions are model extrapolations rather than interpolations. We note that the SED translation is general: applying XClass to a galaxy observed with a different HST filter set requires only specifying the output filter names, and incomplete photometric coverage is handled naturally through degraded fit quality propagated into the feature uncertainties.

3.4. *SED Fit Quality*

The quality of the SED fits varies across the training set. Figure 2 shows the distribution of $\chi^2_{\rm red}$ values for all non-SNR training sources, broken down by class. The median $\chi^2_{\rm red}$ is ∼177, with only ∼10% of sources achieving $\chi^2_{\rm red} < 10$. However, we note that a high $\chi^2_{\rm red}$ does not necessarily indicate an inadequate SED model; it can equally reflect noisy input photometry, underestimated photometric uncertainties (a default of 10% is adopted when per-band errors are unavailable), or incorrect cross-matches in crowded fields. All three effects inflate $\chi^2_{\rm red}$ without implying that the predicted colors are physically unreasonable.

For the subset of sources with well-constrained fits ($\chi^2_{\rm red} < 10$), we verified the internal consistency of the synthetic photometry by comparing predicted HST magnitudes with the input PanSTARRS photometry in the overlapping wavelength range. The predicted F475W and F814W magnitudes reproduce the input PS1 g and i photometry with median offsets of +0.03 and −0.07 mag and scatter (MAD) of 0.05 and 0.06 mag, respectively. The small systematic offsets are consistent with the ∼100 Å wavelength difference between the PS1 and ACS filter pivot wavelengths, confirming that the synthetic photometry machinery is numerically accurate.

To investigate the origin of the high $\chi^2_{\rm red}$ values, we performed an ablation using only PanSTARRS *grizy* photometry as SED input (excluding *Gaia* and 2MASS). The overall median $\chi^2_{\rm red}$ decreased modestly (177.6 to 92.1), but the improvement was class-dependent: HM-STARs improved substantially (median $\chi^2_{\rm red}$ from 209 to 60), while LM-STARs remained poorly fit (median ∼25,000) and X-ray binary fits degraded catastrophically without the NIR anchor needed to constrain the disk fraction. AGN $\chi^2_{\rm red}$ was unchanged (37.2 vs. 36.5), indicating that intrinsic SED diversity — not inter-survey calibration — is the dominant source of fit residu-

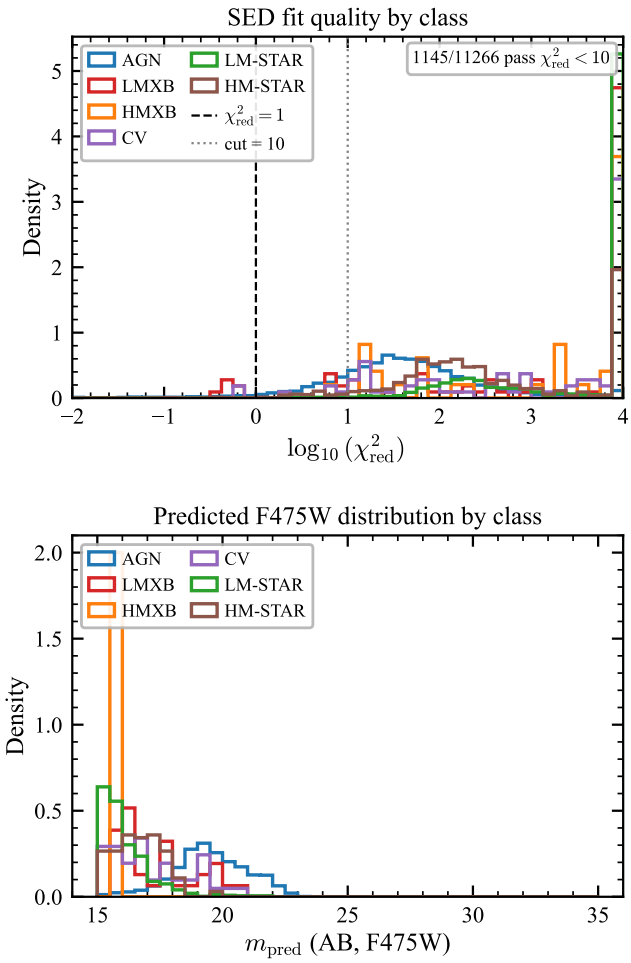

Figure 2. SED fit quality diagnostics. *Top:* Distribution of $\log_{10}(\chi^2_{\rm red})$ for all non-SNR training sources, color-coded by class; the dashed line marks $\chi^2_{\rm red}=10$. *Bottom:* Distribution of predicted F475W magnitudes by class.

als. Classifier performance degraded significantly without 2MASS (balanced accuracy dropped from 0.77 to 0.67), confirming that the NIR bands are essential for XRB classification. We therefore retain the full multi-survey input as the default configuration.

For sources with higher $\chi^2_{\rm red}$, the uncertainty propagation (Equation 4) automatically assigns larger predicted-magnitude errors, appropriately down-weighting the optical/NIR color features for poorly constrained sources. The cross-validation performance reported in Section 5 represents the end-to-end test of whether the full feature set — including sources with high $\chi^2_{\rm red}$ — supports accurate classification.

4. FEATURE ENGINEERING AND CLASSIFICATION

4.1. *Feature Matrix*

The feature matrix comprises 26 columns (for the PHAT filter configuration) organized into six groups (Table 3); the count varies slightly with the target galaxy's HST filter set.

Group 1: X-ray log-fluxes (4 features). The base-10 logarithm of the Chandra flux in the soft (0.5–1.2 keV), medium (1.2–2.0 keV), hard (2.0–7.0 keV), and broad (0.5–7.0 keV) bands.

Group 2: X-ray hardness ratios (4 features). Four ratios are constructed from pairwise flux combinations:

$$\mathrm{HR}_{SM}=\frac{F_{x,S}-F_{x,M}}{F_{x,S}+F_{x,M}},\quad \mathrm{HR}_{MH}=\frac{F_{x,M}-F_{x,H}}{F_{x,M}+F_{x,H}}, \tag{6}$$

$$\mathrm{HR}_{SH}=\frac{F_{x,S}-F_{x,H}}{F_{x,S}+F_{x,H}},\quad \mathrm{HR}_{BM}=\frac{F_{x,B}-F_{x,M}}{F_{x,B}+F_{x,M}}. \tag{7}$$

These ratios are distance-independent and are among the most discriminative features for X-ray source classification (A. H. Prestwich et al. 2003; H. Yang et al. 2022).

Group 3: HST color indices (typically 9 features for PHAT). Consecutive filter colors (e.g., F275W−F336W, F475W−F814W) and wide-baseline colors (e.g., F275W−F814W, F475W−F160W). For training data, SED-translated magnitudes are used, while for application data, real HST magnitudes from the Hubble Source Catalog are used.

Group 4: X-ray to optical flux ratios (2 features). $\log_{10}(F_{x,B}/f_{\nu,\mathrm{F475W}})$ and $\log_{10}(F_{x,B}/f_{\nu,\mathrm{F814W}})$, where f_ν is converted from the AB magnitude. This classical diagnostic separates AGN (high ratios) from stars (low ratios) and is among the most informative features (H. Yang et al. 2022).

Group 5: Match quality and positional features (3 features). The effective Chandra positional uncertainty $\sigma_{\rm eff}$, the normalized separation ν (Section 2.3), and the detection significance. Including ν as a feature allows the classifier to learn that photometry from marginally matched counterparts is less reliable.

Group 6: *Gaia* astrometric features (4 features). $BP-RP$ and $G-RP$ colors, total proper motion $\mu_{\rm tot}$, and a binary flag `is_galactic_parallax` ($\varpi>0.5$ mas). *Gaia* is queried for all sources; foreground stars yield significant parallax, proper motion, and optical colors, while extragalactic sources lack *Gaia* counterparts and their astrometric features are imputed to median values. The contrast between populated and imputed *Gaia* features is itself a strong discriminant

for separating foreground contaminants from sources belonging to the target galaxy.

4.2. *Data Splitting, Imputation, and Class Imbalance*

The training set was split via stratified random sampling into training (72%), validation (18%), and test (10%) subsets, with a fixed random seed for reproducibility. Missing values (arising from incomplete survey coverage, non-detections, or failed SED translations) were imputed with the column median computed on the training set. Any remaining NaN values after imputation were filled with zero. The same procedure is applied identically to application data.

We restrict the training set to sources with at least one non-NaN predicted HST magnitude — either from SED translation (non-SNR classes) or from direct HSC photometry (SNRs). This excludes 890 sources (7% of the CSC-matched sample) for which SED translation failed entirely, avoiding a median-imputation artifact in which sources with all-NaN optical features cluster artificially in feature space. The final training set contains 11,374 sources (Table 1).

To address class imbalance (Table 1), the Random Forest uses `class_weight = ''balanced_subsample''`, recomputing class weights on each bootstrap subsample so that every tree is trained on an effectively balanced dataset. We tested SMOTE oversampling (N. V. Chawla et al. 2002) as a complementary strategy but found that it provides no benefit on the optical baseline: balanced accuracy is marginally higher without SMOTE (0.92 vs. 0.91), and LMXB F1 improves slightly (0.80 vs. 0.76). This is consistent with minority classes already being well-separated in the restricted feature space. SMOTE is therefore disabled by default but remains a configurable option for future applications with different training compositions.

4.3. *Two-Stage Random Forest Architecture*

The classifier operates in two stages, motivated by the difficulty of distinguishing LMXBs from HMXBs at optical wavelengths, where the accretion disk often dominates the donor star's spectral contribution.

Stage 1 (broad classification) assigns each source to one of four groups: **AGN**, **OTHER** (= LMXB + HMXB), **SNR**, and **STAR** (= LM-STAR + HM-STAR).

Stage 2 (X-ray binary subclassification) is trained exclusively on LMXB and HMXB sources and applied only to those that Stage 1 assigns to OTHER. The Stage 2 feature vector augments the original features with Stage 1 class probabilities:

$$\mathbf{x}_2 = [\,\mathbf{x}_1 \parallel p_{\rm AGN}, p_{\rm OTHER}, p_{\rm SNR}, p_{\rm STAR}\,], \qquad (8)$$

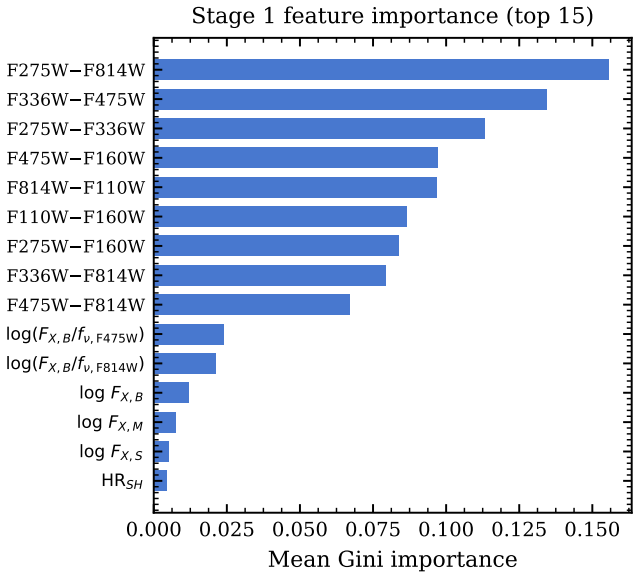

Figure 3. Gini impurity-based feature importance for the top 15 features in the Stage 1 Random Forest (optical baseline), ranked by mean decrease in impurity. Error bars show the 1σ variation across the 600 trees in the ensemble. HST color features exhibit large tree-to-tree variance ($\sigma > \mu$ in some cases) because correlated colors share importance across trees depending on the bootstrap sample; X-ray features (hardness ratios, log fluxes) show tighter error bars because they are less mutually correlated.

where $\parallel$ denotes concatenation. This stacked design allows Stage 2 to exploit Stage 1's calibrated uncertainty as additional discriminative signal.

Both stages use the Random Forest algorithm (L. Breiman 2001) as implemented in `scikit-learn` (F. Pedregosa et al. 2011), configured with 600 trees, unbounded depth, $\sqrt{p}$ features per split, minimum 4 samples to split and 2 per leaf, and balanced subsample class weighting. An XGBoost alternative (T. Chen & C. Guestrin 2016) is also implemented but is not used as the default.

4.4. *Classification Confidence*

Each source receives a confidence score equal to the maximum class probability from the deciding stage. Sources with confidence ≥ 0.5 are flagged as "classified"; those below are flagged as "uncertain."

5. PIPELINE PERFORMANCE

We evaluate classification performance using stratified 5-fold cross-validation. In each fold, a fresh Random Forest was trained on four-fifths of the data and evaluated on the held-out fifth; class proportions are preserved via stratification and results are reported as mean $\pm$ standard deviation across folds. We report

Table 3. Feature matrix for the XClass classifier (PHAT configuration). The 26 features are organized into six groups. The number of HST color features (Group 3) depends on the target filter set; the count shown here is for the six-filter PHAT configuration.

Group	Feature	Description	Source
4*1. X-ray fluxes	$\log F_{x,S}$	Log soft-band flux (0.5–1.2 keV)	CSC 2.0
	$\log F_{x,M}$	Log medium-band flux (1.2–2.0 keV)	CSC 2.0
	$\log F_{x,H}$	Log hard-band flux (2.0–7.0 keV)	CSC 2.0
	$\log F_{x,B}$	Log broad-band flux (0.5–7.0 keV)	CSC 2.0
4*2. Hardness ratios	HR_{SM}	Soft–medium hardness ratio	CSC 2.0
	HR_{MH}	Medium–hard hardness ratio	CSC 2.0
	HR_{SH}	Soft–hard hardness ratio	CSC 2.0
	HR_{BM}	Broad–medium hardness ratio	CSC 2.0
4*3. HST colors	F275W−F336W, ...	Consecutive filter colors	SED trans. / HSC
	F275W−F814W, ...	Wide-baseline colors	SED trans. / HSC
	(9 colors total for PHAT; varies by target)		
2*4. F_X/f_{opt} ratios	$\log(F_{x,B}/f_{\nu,\mathrm{F475W}})$	X-ray to optical ratio (blue)	CSC + SED/HSC
	$\log(F_{x,B}/f_{\nu,\mathrm{F814W}})$	X-ray to optical ratio (red)	CSC + SED/HSC
3*5. Positional	σ_{eff}	Chandra positional uncertainty	CSC 2.0
	ν	Normalized separation	Cross-match
	Significance	Chandra detection significance	CSC 2.0
4*6. *Gaia*	$BP - RP$	*Gaia* optical color	*Gaia* DR3
	$G - RP$	*Gaia* optical color	*Gaia* DR3
	μ_{tot}	Total proper motion (mas yr^{-1})	*Gaia* DR3
	is_galactic_parallax	Parallax > 0.5 mas flag	*Gaia* DR3

four metrics: *accuracy* (fraction correct), *balanced accuracy* (mean per-class recall; K. H. Brodersen et al. 2010), *macro F1* (unweighted mean of per-class F1 scores), and the *Matthews correlation coefficient* (MCC; B. Matthews 1975; D. Chicco & G. Jurman 2020), which is robust to class imbalance. All metrics are reported both for all classifications and for the confident subset (max class probability ≥ 0.5; Section 4.4).

5.1. *Overall Performance*

Table 4 summarizes cross-validation metrics for Stage 1 (four-class) and the full two-stage pipeline (seven-class output), with and without the confidence cut.

5.2. *Confusion Matrices*

Figure 4 presents row-normalized (recall) confusion matrices for Stage 1 and the full pipeline, before and after the confidence cut.

Stage 1 achieves near-perfect separation, with the only off-diagonal entry being a 2% OTHER→STAR leakage. In the full pipeline, LMXB remains the most challenging class (F1 = 0.77), with 27% of LMXBs confused with CV. We find that SNR classification is now robust (F1 = 1.00), confirming that SNRs are cleanly separable when they have genuine optical photometry from the HSC rather than median-imputed features.

5.3. *Per-Class Performance*

Figure 5 shows per-class precision, recall, and F1 from the 5-fold cross-validation. Because Stage 1 maps CVs into the OTHER group and Stage 2 resolves only LMXB vs. HMXB, CV is never a predicted class in the full pipeline and is excluded from the figure. LM-STAR and HM-STAR are evaluated as a single STAR class, consistent with the Stage 1 output.

We find that AGN (F1 = 1.00), STAR (F1 = 1.00), and SNR (F1 = 1.00) are all perfectly or near-perfectly classified. AGN occupy a distinct region of high F_X/f_{opt} and relatively hard X-ray spectra (H. Yang et al. 2022); stars benefit from distinctive *Gaia* astrometric signatures (parallax, proper motion); and SNRs are cleanly separated by their real HSC optical photometry combined with soft X-ray spectra. HMXBs achieve precision of 1.00 and recall of 0.90, yielding F1 = 0.95. LMXBs (precision 0.81, recall 0.73, F1 = 0.77) show the lowest performance: 23 of 86 LMXBs are misclassified, consistent with the small training sample and optical overlap

Table 4. Cross-validation performance metrics from stratified 5-fold out-of-fold evaluation on the optical baseline (11,374 sources, no SMOTE). "All" includes every source; "Confident" restricts to sources with max class probability ≥ 0.5. Values are means across 5 folds ($\pm 1\sigma$).

Metric	Stage 1 (4-class)		Full Pipeline (6-class)	
	All	Confident	All	Confident
Accuracy	0.999 ± 0.001	0.999 ± 0.001	0.996 ± 0.001	0.996 ± 0.001
Balanced Accuracy	0.994 ± 0.003	0.994 ± 0.003	0.907 ± 0.015	0.907 ± 0.015
Macro F1	0.995 ± 0.003	0.995 ± 0.003	0.914 ± 0.012	0.914 ± 0.012
MCC	0.999 ± 0.001	0.999 ± 0.001	0.992 ± 0.002	0.992 ± 0.002
Completeness	1.000	1.000	1.000	1.000

NOTE—All sources exceed the 0.5 confidence threshold, so the "All" and "Confident" columns are identical. The full pipeline uses six evaluation classes (AGN, STAR, SNR, LMXB, HMXB, CV); CV is mapped to OTHER by Stage 1 and is not resolved by Stage 2.

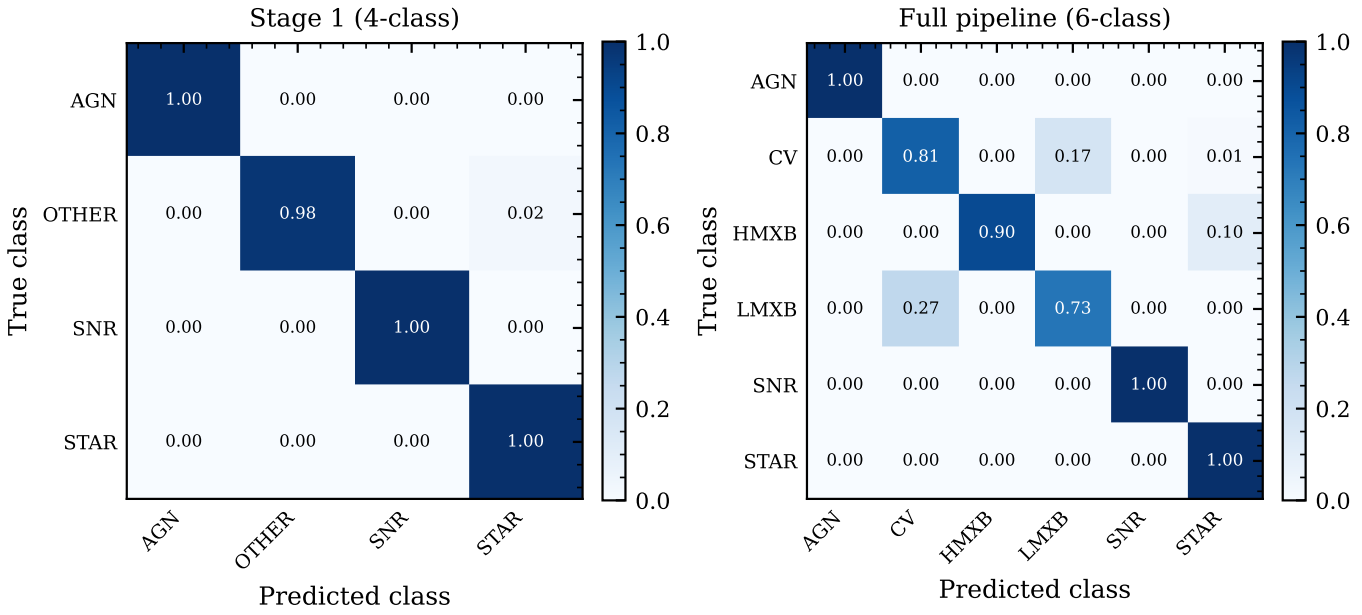

Figure 4. Row-normalized (recall) confusion matrices from 5-fold cross-validation on the optical baseline (11,374 sources). *Left:* Stage 1 (AGN, OTHER, SNR, STAR). *Right:* full two-stage pipeline (AGN, STAR, SNR, LMXB, HMXB). Darker shading indicates higher recall; source counts per class are shown along the vertical axis.

with other compact accreting objects. The fold-level stability is high: AGN, STAR, and SNR achieve F1 = 1.000 in all five folds, while HMXB shows F1 = 0.987 ± 0.027 and LMXB F1 = 0.994 ± 0.012. The fold-to-fold variance for the minority classes is driven by small test-set sizes ($\sim$8 HMXBs and $\sim$17 LMXBs per fold) rather than systematic instability.

5.4. *Confidence Threshold*

Figure 6 shows the distribution of max class probability across all cross-validated sources, separately for correct and incorrect classifications (log-scale y-axis). All sources exceed the 0.5 confidence threshold, with 98.2% falling in the highest bin ($p \geq 0.9$). However, we note that incorrect classifications persist at intermediate confidence levels — the classifier can be confidently wrong, particularly for LMXB sources with optical overlap with

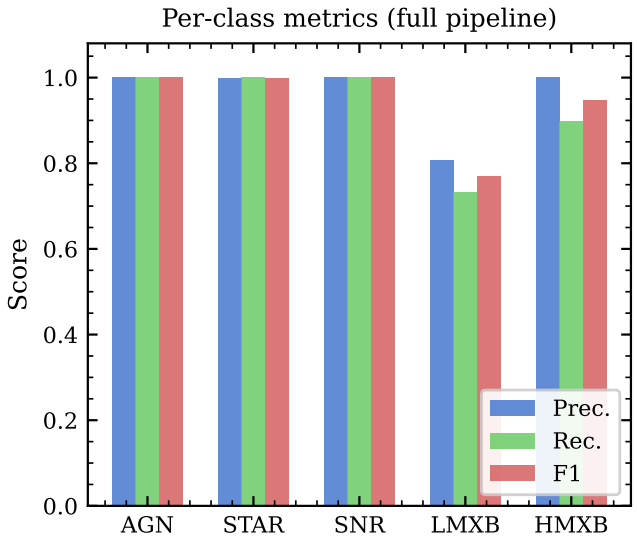

Figure 5. Per-class precision (blue), recall (green), and F1 (orange) from 5-fold cross-validation on the optical baseline (11,374 sources). Classes are ordered by training-set size. LM-STAR and HM-STAR are merged into STAR. CV is excluded because it maps to OTHER in Stage 1 and is not resolved by Stage 2. AGN, STAR, and SNR achieve perfect or near-perfect scores; LMXB (F1 = 0.77) is the only class below 0.9, driven by optical overlap with other compact accreting objects.

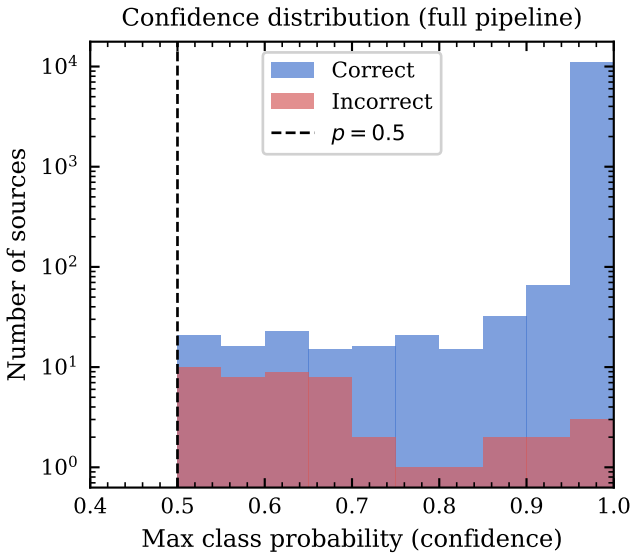

Figure 6. Distribution of max class probability from 5-fold cross-validation on the optical baseline, shown separately for correct (blue) and incorrect (red) classifications on a log-scale y-axis. The dashed line marks the confidence threshold at 0.5. Incorrect classifications persist at intermediate confidence levels.

other compact objects. This is an inherent limitation of the Random Forest vote-fraction probability. A reliability diagram (Figure 7) shows that the classifier is well-calibrated overall, with an expected calibration error (ECE) of 0.002 and a Brier score of 0.003. The classifier is slightly underconfident across all bins (empirical accuracy exceeds stated confidence), which is conservative and well-suited for the intended application. The maximum calibration error (0.17) occurs in the $[0.7, 0.8)$ bin, where 40 sources reside. The full probability vector is provided for every source, enabling users to select their own thresholds.

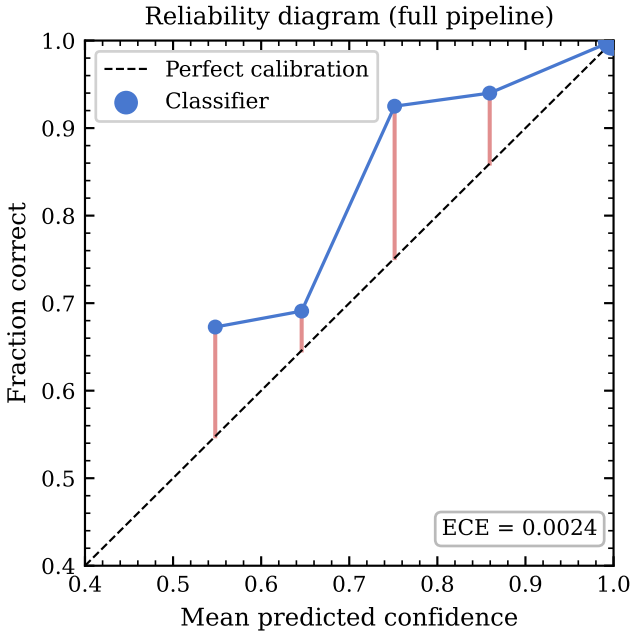

Figure 7. Reliability diagram from 5-fold cross-validation on the optical baseline. The calibration curve (solid) shows the empirical accuracy as a function of predicted confidence; the diagonal represents perfect calibration. The histogram shows the number of sources per confidence bin. The expected calibration error (ECE = 0.002) indicates excellent overall calibration; the classifier is slightly underconfident across all bins.

6. DISCUSSION

6.1. *Assumptions and Limitations*

The SED translation rests on several assumptions that should be taken into account when interpreting classification results.

SED model adequacy. The adopted models (Pickles atlas, AGN composite, two-component star+disk) capture dominant spectral characteristics but do not account for intra-class diversity. The AGN composite (D. E. Vanden Berk et al. 2001) does not represent the range from obscured Seyferts to LINERs, and the two-component model uses fixed donor types (K5 for LMXBs, B2 for HMXBs) whereas real donors span

a range of spectral types and accretion states. These limitations are partially absorbed by the fit uncertainty ($\sigma_{\rm base}$), but systematic color biases may persist for atypical sources.

UV extrapolation. The bluest ground-based constraint (PanSTARRS g at ∼4800 Å) provides no coverage of the UV wavelengths probed by F275W and F336W; thus, UV predictions are model extrapolations. The 0.4 mag systematic (Section 3.3) mitigates this, but UV features remain the least reliably translated — particularly relevant for HMXBs, where UV excess from the massive donor is a key discriminant.

Interstellar extinction. The current implementation does not apply extinction corrections. *Gaia* A_V estimates are recorded but not yet incorporated into the translation. For Galactic training sources, variable line-of-sight extinction alters observed colors (especially in the UV), while for extragalactic targets, internal host-galaxy extinction adds further complication. Incorporating extinction — via dereddening before SED fitting or as an additional model parameter — is a high-priority future improvement.

Source confusion. In crowded galaxy fields, the Chandra positional uncertainty (∼0.″5 for typical on-axis sources) can encompass multiple HST-resolved sources. The current pipeline selects the nearest HSC counterpart and includes the normalized separation ν as a feature, but this is not a substitute for full probabilistic matching. Implementing Bayesian counterpart matching (e.g., NWAY; M. Salvato et al. 2018) is planned for a future version.

Training-set representativeness. The training set is predominantly Galactic and may not fully represent extragalactic X-ray populations. Ultra-luminous X-ray sources and globular-cluster X-ray sources (e.g., M. B. Peacock et al. 2010) are absent and would likely be misclassified as AGN, HMXB, or LMXB. Moreover, the X-ray luminosity function of the training sources reflects Galactic survey sensitivity limits, which may differ from detection thresholds at target-galaxy distances. The HSC concentration index (CI) could in principle be used to identify globular cluster counterparts in future pipeline versions.

Flat-disk approximation. The accretion disk is modeled as a flat ($f_\nu = \text{const}$) spectrum (J. Frank et al. 2002), a reasonable zeroth-order approximation but one that ignores dependences on accretion rate, inclination, and irradiation.

SNR optical features. SNRs bypass SED translation and instead use real HST photometry from the Hubble Source Catalog. Of the 150 SNR training sources, 108 (72%) have HSC detections in at least one PHAT filter; the remaining 42 — primarily diffuse remnants in M33 below the HSC point-source detection threshold — lack optical photometry entirely and are excluded from the training set by the ≥ 1 optical magnitude criterion (Section 4.2). With this restriction, SNR classification achieves perfect recall and precision in cross-validation, confirming that the classifier identifies SNRs from genuine astrophysical signals rather than imputation artifacts.

SMOTE oversampling. We tested the classifier with and without SMOTE on the optical baseline. SMOTE provides no benefit on this dataset: balanced accuracy is 0.92 without SMOTE versus 0.91 with it, and LMXB F1 improves slightly without oversampling (0.80 vs. 0.76). This is consistent with the finding that minority classes are already well-separated in the optical baseline feature space; SMOTE-generated synthetic points in this regime introduce marginal confusion rather than improving coverage. We retain SMOTE as a configurable option for future applications with different training compositions, but it is disabled by default.

Comparison with MUWCLASS. A direct quantitative comparison with MUWCLASS (H. Yang et al. 2022) is complicated by differences in class definitions, feature sets, and validation methodology. However, both pipelines achieve high accuracy for populous classes and find LMXB/HMXB confusion to be the dominant source of classification error.

Class-aware SED model selection. The SED translation uses the known class label to select the physical model (Table 2). To quantify the impact of this choice, we performed an ablation in which all training sources were fit with a single class-blind model (Pickles templates or blackbody for all classes, regardless of label). The class-blind configurations reduced Stage 1 accuracy by 8–10 percentage points and macro F1 by 21–24 percentage points, but still achieved >88% accuracy, indicating that the classifier is not wholly dependent on label-informed SED choices. We note that this design is not circular label leakage but rather the injection of physical domain knowledge: fitting an AGN with an AGN composite and a star with a stellar template produces more physically accurate translated colors than forcing a single model on all source types. Moreover, the SED translation applies only to the training set. At application time, extragalactic sources have real HST photometry directly from the HSC, so no model selection — class-aware or otherwise — is involved in the classification of new sources.

6.2. *Future Improvements*

Several enhancements are planned. Incorporating *Gaia* extinction estimates, and 3D dust maps (G. M. Green et al. 2019) into the SED fitting would improve translated colors, particularly for training sources in the Galactic plane. Expanding the training set — especially the underrepresented LMXB, HMXB, CV, and SNR classes, and potentially incorporating validated extragalactic classifications from M31/M33 as feedback — would directly improve minority-class performance. In addition, adding X-ray variability features from CSC 2.0 would enhance discrimination between transient and persistent sources, and implementing full Bayesian counterpart matching via NWAY (M. Salvato et al. 2018) for both training and application would improve robustness in crowded fields. As the training set grows, deep learning architectures may offer improved feature interactions, the modular pipeline design allows the classifier to be swapped without modifying upstream processing. Finally, the SED translation framework can be adapted to future facilities: *Athena* (K. Nandra et al. 2013), AXIS (R. Mushotzky 2018), and the Einstein Probe (W. Yuan et al. 2022) will detect large numbers of extragalactic X-ray sources, while the Vera C. Rubin Observatory (Ž. Ivezić et al. 2019), the *Roman Space Telescope*, and *Euclid* (R. Laureijs et al. 2011) will provide deep optical/NIR coverage. Adapting XClass to use Rubin/LSST *ugrizy* as the application filter system would enable classification beyond HST's limited field of view.

7. SUMMARY AND CONCLUSIONS

We have presented XClass (X-ray Classifier for extragalactic sources), an end-to-end machine-learning pipeline for classifying extragalactic X-ray point sources into seven astrophysical classes (AGN, LMXBs, HMXBs, CVs, LM-STARs, HM-STARs, SNRs). The main results are as follows.

1. We developed an SED translation procedure that bridges the photometric gap between ground-based training data (PanSTARRS, *Gaia*, 2MASS) and HST application data by fitting class-appropriate physical spectral models and convolving them through HST filter curves, producing synthetic magnitudes in a common feature space. The approach is general and applicable to any HST filter configuration. SED fit quality varies across the training set, but the uncertainty propagation automatically down-weights poorly constrained translated magnitudes; the cross-validation metrics in Section 5 represent the end-to-end performance of the full pipeline (Section 3).

2. We assembled a training dataset from ten VizieR catalogs of Galactic X-ray sources and extragalactic SNR catalogs (M33, NGC 6946, M83), cross-matched to CSC 2.0 detections with photometry from three ground-based surveys (see Table 1).

3. We constructed a 26-feature matrix combining X-ray fluxes and hardness ratios, SED-translated HST colors, X-ray–to–optical flux ratios, positional match-quality metrics, and *Gaia* astrometric properties (Section 4.1).

4. We implemented an asymmetric two-stage Random Forest: Stage 1 separates AGN, OTHER (LMXB+HMXB), SNR, and STAR; Stage 2 resolves OTHER into LMXB or HMXB using an augmented feature vector that includes Stage 1 probabilities (Section 4.3).

5. We evaluated performance via stratified 5-fold cross-validation on the optical baseline (11,374 sources, no SMOTE). Stage 1 achieves 99.9% accuracy; the full pipeline achieves 99.6% accuracy and balanced accuracy of 0.91 (macro F1 = 0.91, ECE = 0.002). AGN, STAR, SNR, and HMXB are classified with F1 $\geq$ 0.95; LMXB confusion (F1 = 0.77) is the dominant remaining source of error (Section 5).

6. The pipeline is implemented as a modular, open-source Python package with centralized configuration, per-source disk caching, and a CLI supporting stage-by-stage execution.

XClass will be applied to M31 and M33 in a forthcoming companion paper. The SED translation framework is adaptable to future facilities, including *Athena*, AXIS, the Einstein Probe, the Vera C. Rubin Observatory, and the *Roman Space Telescope*, supporting classification of ever-larger extragalactic X-ray source samples.

ACKNOWLEDGMENTS

Support for this work was provided by the National Aeronautics and Space Administration (NASA) through Chandra Award Number AR8-19009A issued by the Chandra X-ray Center, which is operated by the Smithsonian Astrophysical Observatory for and on behalf of NASA under contract NAS8-03060, and by NASA under award number 80NSSC22K1119 issued through the Astrophysics Data Analysis Program (ADAP). This research has made use of data obtained from the Chandra Source Catalog, provided by the Chandra X-ray Center (CXC) as part of the Chandra Data Archive. Claude (

Anthropic 2025) was used for assistance with code editing, data visualization, and language refinement.

The HST data used in this paper can be found in MAST: 10.17909/T97P46 (Hubble Source Catalog) and 10.17909/s0zg-jx37 (PanSTARRS DR2).

Facilities: CXO, HST(ACS), HST(WFC3), MAST

Software: astropy (Astropy Collaboration et al. 2013, 2018, 2022), scikit-learn (F. Pedregosa et al. 2011), numpy (C. R. Harris et al. 2020), matplotlib (J. D. Hunter 2007), pandas (W. McKinney 2010), Claude (Anthropic 2025)

REFERENCES

Anthropic. 2025, Claude, https://www.anthropic.com/claude

Arnason, R. M., Barmby, P., & Vulic, N. 2020, MNRAS, 492, 5075, doi: 10.1093/mnras/staa207

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33, doi: 10.1051/0004-6361/201322068

Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123, doi: 10.3847/1538-3881/aabc4f

Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, ApJ, 935, 167, doi: 10.3847/1538-4357/ac7c74

Blair, W. P., Winkler, P. F., & Long, K. S. 2012, ApJS, 203, 8, doi: 10.1088/0067-0049/203/1/8

Brassington, N. J., Fabbiano, G., Zezas, A., et al. 2012, ApJ, 755, 162, doi: 10.1088/0004-637X/755/2/162

Breiman, L. 2001, Machine Learning, 45, 5, doi: 10.1023/A:1010933404324

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. 2010, in Proceedings of the 2010 20th International Conference on Pattern Recognition, ICPR '10 (USA: IEEE Computer Society), 3121–3124, doi: 10.1109/ICPR.2010.764

Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, arXiv e-prints, arXiv:1612.05560, doi: 10.48550/arXiv.1612.05560

Charles, P. A., & Coe, M. J. 2006, in Compact Stellar X-ray Sources, ed. W. H. G. Lewin & M. van der Klis (Cambridge University Press), 215–265, doi: 10.1017/CBO9780511536281.006

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002, Journal of Artificial Intelligence Research, 16, 321, doi: 10.1613/jair.953

Chen, S., Kargaltsev, O., Yang, H., et al. 2023, ApJ, 948, 59, doi: 10.3847/1538-4357/acb3a6

Chen, T., & Guestrin, C. 2016, arXiv e-prints, arXiv:1603.02754, doi: 10.48550/arXiv.1603.02754

Chicco, D., & Jurman, G. 2020, BMC Genomics, 21, 6, doi: 10.1186/s12864-019-6413-7

Dalcanton, J. J., Williams, B. F., Lang, D., et al. 2012, ApJS, 200, 18, doi: 10.1088/0067-0049/200/2/18

Downes, R. A., Webbink, R. F., Shara, M. M., et al. 2001, PASP, 113, 764, doi: 10.1086/320802

Evans, I. N., Primini, F. A., Miller, J. B., et al. 2020, Bulletin of the American Astronomical Society, 52

Evans, I. N., Primini, F. A., Glotfelty, K. J., et al. 2010, ApJS, 189, 37, doi: 10.1088/0067-0049/189/1/37

Fabbiano, G. 2006, ARA&A, 44, 323, doi: 10.1146/annurev.astro.44.051905.092519

Frank, J., King, A., & Raine, D. J. 2002, Accretion Power in Astrophysics, 3rd edn. (Cambridge University Press), doi: 10.1017/CBO9781139164245

Gaia Collaboration, Vallenari, A., Brown, A. G. A., et al. 2023, A&A, 674, A1, doi: 10.1051/0004-6361/202243940

Gaskin, J. A., Swartz, D. A., Vikhlinin, A., et al. 2019, Journal of Astronomical Telescopes, Instruments, and Systems, 5, 021001, doi: 10.1117/1.JATIS.5.2.021001

Gilfanov, M. 2004, MNRAS, 349, 146, doi: 10.1111/j.1365-2966.2004.07473.x

Ginsburg, A., Sipőcz, B. M., Brasseur, C. E., et al. 2019, AJ, 157, 98, doi: 10.3847/1538-3881/aafc33

Green, G. M., Schlafly, E., Zucker, C., Speagle, J. S., & Finkbeiner, D. 2019, ApJ, 887, 93, doi: 10.3847/1538-4357/ab5362

Grimm, H.-J., Gilfanov, M., & Sunyaev, R. 2003, MNRAS, 339, 793, doi: 10.1046/j.1365-8711.2003.06224.x

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357, doi: 10.1038/s41586-020-2649-2

Hunter, J. D. 2007, Computing in Science & Engineering, 9, 90, doi: 10.1109/MCSE.2007.55

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111, doi: 10.3847/1538-4357/ab042c

Kaplan, D. L., Gaensler, B. M., Kulkarni, S. R., & Slane, P. O. 2006, ApJS, 163, 344, doi: 10.1086/501441

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints, arXiv:1110.3193, doi: 10.48550/arXiv.1110.3193

Liu, Q. Z., van Paradijs, J., & van den Heuvel, E. P. J. 2006, A&A, 455, 1165, doi: 10.1051/0004-6361:20064987

Liu, Q. Z., van Paradijs, J., & van den Heuvel, E. P. J. 2007, A&A, 469, 807, doi: 10.1051/0004-6361:20077303

Lo, K. K., Farrell, S., Murphy, T., & Gaensler, B. M. 2014, ApJ, 786, 20, doi: 10.1088/0004-637X/786/1/20

Long, K. S., Winkler, P. F., & Blair, W. P. 2019, ApJ, 875, 85, doi: 10.3847/1538-4357/ab0d94

Long, K. S., Blair, W. P., Winkler, P. F., et al. 2010, ApJS, 187, 495, doi: 10.1088/0067-0049/187/2/495

Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94, doi: 10.3847/1538-3881/aa784d

Matthews, B. 1975, Biochimica et Biophysica Acta (BBA) - Protein Structure, 405, 442, doi: https://doi.org/10.1016/0005-2795(75)90109-9

McKinney, W. 2010, in Proceedings of the 9th Python in Science Conference, ed. S. van der Walt & J. Millman, 51–56

Misanovic, Z., Kargaltsev, O., & Pavlov, G. G. 2010, ApJ, 725, 931, doi: 10.1088/0004-637X/725/1/931

Mushotzky, R. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10699, Space Telescopes and Instrumentation 2018: Ultraviolet to Gamma Ray, ed. J.-W. A. den Herder, S. Nikzad, & K. Nakazawa, 1069929, doi: 10.1117/12.2310003

Nandra, K., Barret, D., Barcons, X., et al. 2013, arXiv e-prints, arXiv:1306.2307, doi: 10.48550/arXiv.1306.2307

Ochsenbein, F., Bauer, P., & Marcout, J. 2000, A&AS, 143, 23, doi: 10.1051/aas:2000169

Peacock, M. B., Maccarone, T. J., Knigge, C., et al. 2010, MNRAS, 402, 803, doi: 10.1111/j.1365-2966.2009.15952.x

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825. http://jmlr.org/papers/v12/pedregosa11a.html

Pickles, A. J. 1998, PASP, 110, 863, doi: 10.1086/316197

Prestwich, A. H., Irwin, J. A., Kilgard, R. E., et al. 2003, ApJ, 595, 719, doi: 10.1086/377366

Rangelov, B., Yang, H., Williams, B., et al. 2024, ApJ, 961, 26, doi: 10.3847/1538-4357/ad09da

Ritter, H., & Kolb, U. 2003, A&A, 404, 301, doi: 10.1051/0004-6361:20030330

Rodrigo, C., & Solano, E. 2020, in XIV.0 Scientific Meeting (virtual) of the Spanish Astronomical Society, 182

Salvato, M., Buchner, J., Budavári, T., et al. 2018, MNRAS, 473, 4937, doi: 10.1093/mnras/stx2651

Sirianni, M., Jee, M. J., Benítez, N., et al. 2005, PASP, 117, 1049, doi: 10.1086/444553

Skiff, B. A. 2014, VizieR Online Data Catalog: Catalogue of Stellar Spectral Classifications (Skiff, 2009-2014),, VizieR On-line Data Catalog: B/mk. Originally published in: 2014yCat....1.2023S

Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163, doi: 10.1086/498708

Stiele, H., Pietsch, W., Haberl, F., et al. 2011, A&A, 534, A55, doi: 10.1051/0004-6361/201015270

Tranin, H., Godet, O., Webb, N., & Primorac, D. 2022, A&A, 657, A138, doi: 10.1051/0004-6361/202141259

Tüllmann, R., Gaetz, T. J., Plucinsky, P. P., et al. 2011, ApJS, 193, 31, doi: 10.1088/0067-0049/193/2/31

van der Hucht, K. A. 2001, NewAR, 45, 135, doi: 10.1016/S1387-6473(00)00112-3

van der Hucht, K. A. 2006, A&A, 458, 453, doi: 10.1051/0004-6361:20065819

Vanden Berk, D. E., Richards, G. T., Bauer, A., et al. 2001, AJ, 122, 549, doi: 10.1086/321167

Véron-Cetty, M.-P., & Véron, P. 2010, A&A, 518, A10, doi: 10.1051/0004-6361/201014188

Williams, B. F., Lazzarini, M., Plucinsky, P. P., et al. 2018, ApJS, 239, 13, doi: 10.3847/1538-4365/aae37d

Yang, H., Hare, J., Kargaltsev, O., et al. 2022, ApJ, 941, 104, doi: 10.3847/1538-4357/ac952b

Yuan, W., Zhang, C., Chen, Y., & Ling, Z. 2022, in Handbook of X-ray and Gamma-ray Astrophysics, ed. C. Bambi & A. Sangangelo, 86, doi: 10.1007/978-981-16-4544-0_151-1