Investigation of emission and variability evolution of low mass X-ray binaries using machine learning techniques

M.Sc. Thesis

Kaivan S. S.



Department of Astronomy, Astrophysics and Space Engineering
INDIAN INSTITUTE OF TECHNOLOGY INDORE
May 18, 2026

Investigation of emission and variability evolution of low mass X-ray binaries using machine learning techniques

M.Sc.Thesis

Submitted in partial fulfillment of the requirements for the awards of the degree
of
Master of Science

by

Kaivan S. S.



Department of Astronomy, Astrophysics and Space Engineering

INDIAN INSTITUTE OF TECHNOLOGY INDORE

May 18, 2026

Acknowledgements

I would like to sincerely thank my supervisor, **Dr. Manoneeta Chakraborty**, for her constant guidance, support, and encouragement throughout this work. Her patience, suggestions, and approachable nature greatly contributed to my learning and academic growth. I am truly grateful for the opportunities and research perspective I gained under her supervision.

I am thankful to my lab seniors, Biki bhai, Aromal, and Jibin, for their constant support, discussions, motivation, and for generously sharing their experiences and knowledge, all of which greatly contributed to my understanding of the research domain and learning throughout this work. I would like to thank Yash bhai, Mohit bhai, and Vishrut bhai for their valuable suggestions and insights that improved the thesis.

I also extend my appreciation to my lab mate, Nityananda, for several meaningful academic and personal discussions that provided fresh perspectives throughout this journey. I sincerely thank Sangeetha for her encouragement throughout the work. I am thankful to my colleagues Mayukh, Kartik, Manpreet, Aditya, Sangeeta, Kanishka, and Vishal for making this journey truly memorable.

Last but not least, I would like to thank my **parents** and **sister** for their constant support and encouragement, without which this work would not have been possible.

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Investigation of emission and variability evolution of low mass X-ray binaries using machine learning techniques** in the partial fulfillment of the requirements for the award of the degree of **MASTER OF SCIENCE** and submitted in the **Department of Astronomy, Astrophysics and Space Engineering, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from June 2025 to May 2026 under the supervision of Dr. Manoneeta Chakraborty, Associate Professor.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

Kaivan 08/05/26

Signature of the student with date
Kaivan Sanjay Shah

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Manoneeta Chakraborty

8/5/2026

Signature of the Supervisor of

M.Sc. thesis (with date)

Dr. Manoneeta Chakraborty

Kaivan Sanjay Shah has successfully given his M.Sc. Oral Examination held on **06/05/2026.**

Manoneeta Chakraborty

Signature(s) of Supervisor(s) of MSc thesis
Date: 8/5/2026

Manoneeta Chakraborty

Programme Coordinator, M.Sc.
Date: 8/5/2026

Manoneeta Chakraborty

Convenor, DPGC
Date: 8/5/2026

HoD, DAASE
Date: 08/05/2026

Abstract

During an outburst, neutron star low-mass X-ray binaries (NS LMXBs) follow a specific evolutionary track in their color–color and hardness–intensity diagrams. These tracks categorize NS LMXBs into subclasses, suggesting that sources within the same subclass may share similar accretion behavior. As the outburst progresses, variations in accretion rate and the disk–corona structure lead to changes in the source's spectral state. Transitions in spectral state influence the source's variability, providing observational signatures of changing physical conditions. Therefore, classifying the sources into their subclasses and corresponding spectral states is important for understanding the accretion process and system geometry. While machine learning approaches have been used for such studies, NS LMXBs remain relatively unexplored in this context, particularly with AstroSat data.

To address this challenge, we employ machine learning (ML) techniques that focus on both source-class identification and spectral-state identification for the Atoll subclass. We used archival data from the LAXPC instrument on board the *AstroSat* satellite, comprising 35 sources with a total of 133 observations. A six-dimensional feature space was constructed using spectral hardness parameters (hard color and soft color), intensity, and timing-variability features (power colors PC1 and PC2, and fractional RMS amplitude).

For source-class classification (Atoll, Type-Z, Accreting Millisecond X-ray Pulsar, and Atoll + Type-Z) and spectral-state classification within Atoll sources, four supervised machine learning algorithms were employed and compared. The K-Nearest Neighbor classifier performed best with the overall accuracy of $84.9 \pm 0.2\%$.

For spectral-state classification within Atoll sources, all models show strong performance, with SVM achieving the highest accuracy of $93.6 \pm 0.6\%$. The timing-variability features, particularly PC2, were found to be more important for the classification. We also used an unsupervised Gaussian Mixture Model (GMM), which identifies clusters corresponding to spectral states, with one cluster indicating a transitional region.

This study shows that machine learning methods are effective for classifying NS LMXBs and can reproduce the expected source and spectral state identifications, indicating that the chosen feature space captures the relevant properties of the sources. This approach also enables quicker identification of source properties in newly detected outbursts, allowing timely multi-wavelength follow-up observations to be planned more efficiently.

Contents

List of Figures

List of Tables

1 Introduction **1**

1.1 Compact Objects . . . 1

1.2 X-ray Binaries . . . 2

1.2.1 Type of X-ray Binaries . . . 2

1.3 Classification of Low-Mass X-ray binary . . . 4

1.4 Outbursts in Neutron Star Low-Mass X-ray Binaries . . . 5

1.4.1 Evolution of a source during an outburst . . . 5

1.4.2 Classification of Neutron Star Low-Mass X-ray Binaries . . . 6

1.5 Motivation . . . 9

1.5.1 Review of related works . . . 10

1.5.2 Objectives . . . 12

2 Instrument **13**

2.1 LAXPC instrument . . . 14

3 Observations and Data Preparations **16**

3.1 Data Processing . . . 16

3.1.1 Sample Selection . . . 16

3.1.2 Data Reduction . . . 20

3.2 Feature Extraction . . . 21

3.2.1 Hardness Ratios . . . 21

3.2.2 Intensity . . . 22

3.3 Power Color and Fractional RMS Amplitude . . . 24

3.4 Dataset Labeling . . . 26

3.4.1 Source-class Classification . . . 26

3.4.2 Spectral State based Classification . . . 28

3.5 Feature Visualization . . . 31
3.5.1 Hardness-Intensity Diagram and Color-Color Diagram . . . 31
3.5.2 Color-Color-Intensity Diagram . . . 35
3.5.3 Power Color Diagram (PCD) . . . 38

4 Methodology **39**
4.1 Classification Algorithm . . . 39
4.1.1 K-Nearest Neighbors . . . 39
4.1.2 Support Vector Machine . . . 40
4.1.3 Random Forest (RF) . . . 42
4.1.4 XGBoost (eXtreme Gradient Boosting) . . . 43
4.2 Training dataset preparation . . . 43
4.3 Cross Validation and Model Hyperparameter Tuning . . . 45
4.3.1 Source-class Classification: K-Nearest Neighbors . . . 46
4.3.2 State Classification for Atolls: Support Vector Machine . . . 48
4.4 Unsupervised Machine Learning Method . . . 49
4.4.1 Clustering Algorithm . . . 50
4.4.2 Hyperparameter Tuning . . . 51

5 Results and Discussions **53**
5.1 Source-class Classification . . . 53
5.2 State Classification for Atoll sources . . . 58
5.2.1 Feature Importance . . . 61
5.2.2 Model's behavior across state-transition regions . . . 62
5.3 Comparing Algorithms . . . 64
5.3.1 For Source-Class classification . . . 64
5.3.2 For State classification . . . 64
5.3.3 Results from Unsupervised Learning: Gaussian Mixture Model . . . 66

6 Conclusion **68**

7 Future directions **71**

List of Figures

1.1 An artistic impression of an X-ray binary system. [Credits: ESA] 2
1.2 Schematic diagram illustrating the stages of Roche Lob overflow. [Credit: Swinburne University of Technology] 3
1.3 HID and CCD representation of a typical Atoll type LMXB. Image reference: Klein-Wolt & Van Der Klis (2008) 7
1.4 CCD plot of Type-Z source using RXTE data. Image reference: Chaty (2022) . 8

2.1 AstroSat satellite carrying five payloads 13
2.2 Schematic showing the three identical LAXPC units onboard *AstroSat*. 14
2.3 Comparison of the effective area of LAXPC and the other instruments of ASTROSAT with all the other dedicated astronomical X-ray missions 15

3.1 Number of data points per source used in the analysis. Each data point corresponds to a 512 s segment of the light curve, from which the hard color, soft color, and intensity are extracted. 24
3.2 Schematic representation of spectral states in the power color diagram (PC1–PC2 plane). For the representation of the ellipse, the center point was chosen as [4.519,0.453] by the authors. [Image reference: Heil et al.] 26
3.3 Hardness–Intensity Diagram (HID; left) and Colour–Colour Diagram (CCD; right) for selected sources. A representative error bar (black) is shown for a single data point for clarity; error bars for all points are omitted to avoid visual clutter. 33
3.4 Combined HID and CCD showing all sources in gray. The highlighted source class is marked to indicate its position within the overall feature dataset. 35
3.5 Three-dimensional Color–Color–Intensity (CCI) diagram visualized from different viewing orientations. The CCI is defined in terms of hard color, soft color, and normalized intensity along three axes. 36
3.6 Color–Color–Intensity (CCI) diagrams for different source classes (Atoll, Z, Pulsar, and Atoll+Z). The highlighted points correspond to the selected class, while the grey points represent all other data. 37

3.7 Power colour diagram plotted in $\log_{10}$ scale for each power color value. Data points are color-coded by spectral state, while grey points represent unlabeled data. 38

4.1 A Flowchart of the K-nearest neighbor algorithm is shown. Once the K value is chosen, the algorithm calculates distances from a single test point to all points in the training dataset, then selects the class with the majority count among the nearest neighbors. The process is repeated for multiple test data points. 40

4.2 The figure provides a schematic for the SVM algorithm for the margin width maximization. Here, the width is maximized so that data points from two classes (represented by plus and minus signs in the figure) are well separated. [Image reference: `https://uk.mathworks.com/discovery/support-vector-machine.html`] 41

4.3 Schematic representation of the Random Forest (RF) algorithm, illustrating the bagging approach used to train multiple decision trees. The final prediction is obtained through majority voting among the individual tree outputs. [Image reference: `https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/`] 42

4.4 Distribution of dataset labels: The top panel (previous page) displays class labels, while this panel illustrates atoll state labels. 44

4.5 The plot of K values vs the F1 score is given, where the scores are an aggregate of random 50 seed values. For K = 10, the model performs better compared to other values of K. 47

4.6 Adjusted Rand Index (ARI) as a function of the number of Gaussian components used in the GMM. 52

5.1 Figure shows class-wise classification performance scores (precision, recall, and F1-score, respectively) for source-class-based classification. 54

5.2 Confusion matrix for the KNN method used for source class classification. . . . 54

5.3 Accuracy of source classification under the Leave-One-Source-Out method. . . 56

5.4 Average signal-to-noise ratio (SNR) distribution across different source classes, along with the corresponding classification accuracy. Pulsar sources exhibit comparatively lower SNR values, which correlate with reduced classification performance. 57

5.5 Figure shows class-wise classification performance scores (precision, recall, and F1-score, respectively) for state-classification. 58

5.6 Confusion matrix for the SVM method used for state classification. 59

5.7 Figure shows the accuracy of state classification under the Leave-One-Source-Out method. 60
5.8 Feature importance scores for the input features, computed as the decrease in accuracy when that feature is excluded. 61
5.9 Figure provides the HID and PCD projection, here the color bar represents the classification probability for each sample. 62
5.10 The figure shows SVM predictions on the unlabeled subset of the feature space. The SVM classifier is trained on the entire training dataset. 63
5.11 Plot of performance metric vs different models is shown for source-class-based classification. The algorithms on the x-axis are (left to right) K-nearest Neighbor, Support Vector Machine, Random Forest, and eXtreme Gradient Boosting. 64
5.12 Plot of performance metric vs different models is shown for spectral state classification of atoll sources. The algorithms on the x-axis are (left to right) K-nearest Neighbor, Support Vector Machine, Random Forest, and eXtreme Gradient Boosting. 65
5.13 Overlap matrices showing the relationship between GMM clusters and labeled spectral states. The upper panel shows cluster-wise state composition, while the lower panel gives the distribution of each state across clusters. 67

List of Tables

3.1 Observation details of selected X-ray binary sources 17
3.2 List of selected neutron star low-mass X-ray binary sources along with their distances and normalization factor $(d/5)^2$. 23
3.3 Classification of X-ray binary sources used in this work 27
3.4 List of observations along with inferred spectral states and corresponding references. 29

4.1 Grid search results for SVM hyperparameters for rbf kernel. 49

5.1 Overall classification performance metrics for the source-class classification using the KNN model. Errors represent the standard deviation (1σ) across all cross-validation folds. 53
5.2 Summary of misclassified sources in source-class identification using Leave-One-Source-Out (LOSO) evaluation. 56
5.3 Overall classification performance metrics for the state classification using the SVM model. Errors represent the standard deviation (1σ) across all cross-validation folds. 59

Chapter 1

Introduction

1.1 Compact Objects

The majority of stars observed in our galaxy, including the Sun, are found in the main sequence phase of their evolution. During this stage, the stellar structure is maintained by a balance between gravitational contraction and the outward thermal pressure generated by the nuclear fusion in the core. Stars spend most of their lifetime on the main sequence, converting hydrogen into helium through nuclear fusion. When it runs out of this fuel, the star cools and collapses under its own gravity. The collapse heats the core, triggering successive stages of fusion; however, how far this process continues depends on the mass of the star (low and intermediate mass stars can fuse only up to carbon or oxygen, while very massive stars can continue fusion up to heavier elements, eventually producing iron in the core). Once iron is formed, nuclear reactions cease, and the core can no longer sustain the outward pressure, leading to further gravitational contraction. This triggers fusion in the surrounding shell, and the energy released from the shell burning causes the outer layers to expand. The layers are eventually expelled, leaving behind the dense stellar remnant known as a compact object. The compact object can be a White Dwarf (supported by electron degenerate pressure), a Neutron star (supported by neutron degenerate pressure), or a Black Hole, depending on the mass of the progenitor.

For progenitors with masses in the range $8\,M_\odot \lesssim M \lesssim 20\,M_\odot$, the collapse of the stellar core results in the formation of a neutron star. These are highly compact objects with roughly 1.4 $M_\odot$ of mass concentrated within a radius of ~10 km (Cerda-Duran & Elias-Rosa, 2018). Whereas massive progenitors with mass M > $20M_\odot$, collapse further to form a

black hole (Fryer, 1999).

Such compact objects serve as natural laboratories for studying matter under high gravitational potential, temperature, and density.

1.2 X-ray Binaries

Binary systems consist of two massive objects that orbit a common center of mass. In an X-ray binary system, one of the components is a compact object (black hole, neutron star, or white dwarf), while the other is a normal star (usually a main-sequence star). Such systems originate from binary star pairs in which the more massive component evolves faster and ends up leaving a compact object as a supernova remnant. If the system remains gravitationally bound after the explosion, the compact object can begin to accrete matter from its companion (Hills, 1980). As the infalling material spirals inwards, it loses gravitational potential energy, which is converted to heat and radiation (Frank et al., 2002). This process raises the temperature of accreting gas to the order of 10^7K, resulting in strong emissions primarily in the X-ray band.

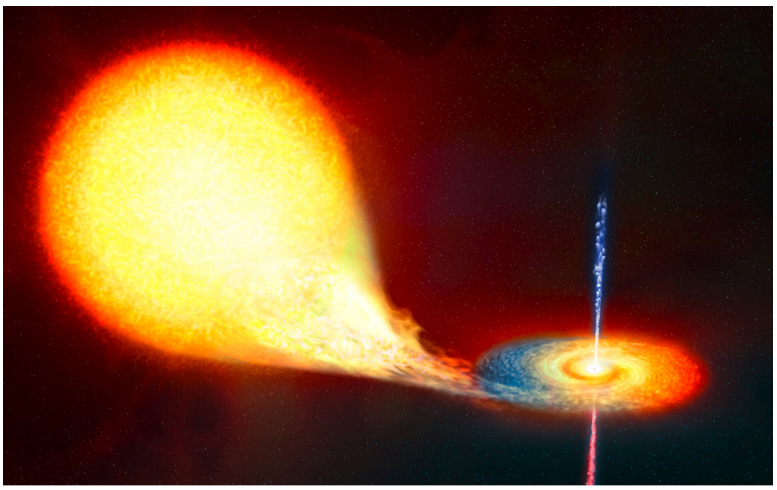


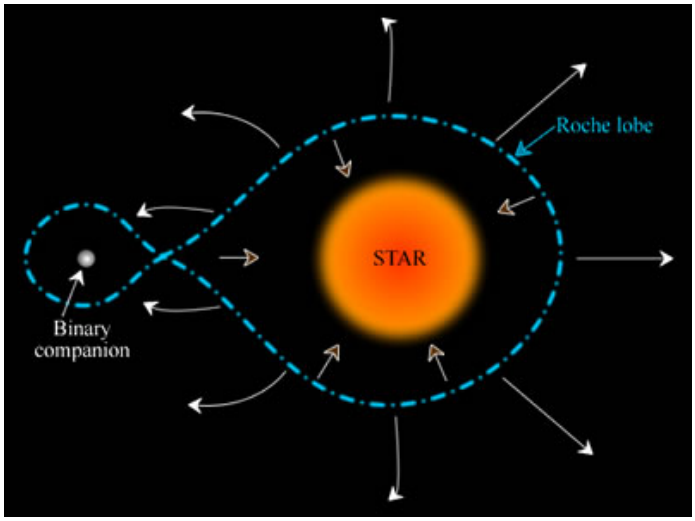
Figure 1.1: An artistic impression of an X-ray binary system. [Credits: ESA]

1.2.1 Type of X-ray Binaries

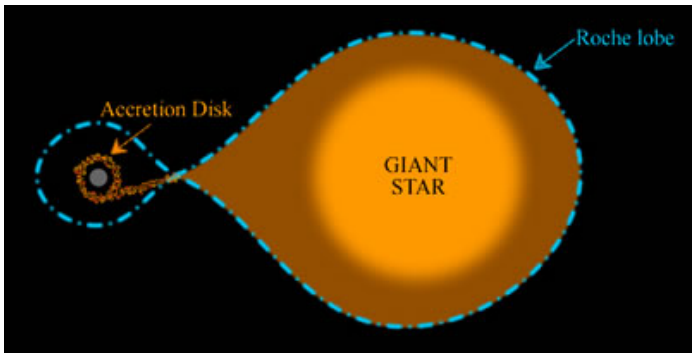
On the basis of the mass of the companion star, X-ray binaries (XRBs) can be classified into two categories: Low Mass X-ray binaries (LMXBs) and High Mass X-ray binaries (HMXBs). In LMXBs, the companion star has a mass of M $\lesssim 1M_{\odot}$, whereas in the HMXBs, the companion star has a mass of M $\gtrsim 10M_{\odot}$ (Lewin et al., 1995; Remillard & McClintock,

2006). Due to this difference in the mass of the companion star, both systems follow different accretion processes.

In a binary system, each star has its own region of gravitational dominance within which any orbiting matter remains bound to the star. These regions are called Roche Lobes (a tear-drop shaped region), and in a star–companion system, the two Roche Lobes touch at a point known as the inner Lagrangian point (L_1). As the companion star ages, during its red giant phase, it expands, and its outer layers may start filling the Roche Lobe. Once this occurs, matter gets accreted on the compact object as it flows through the inner Lagrangian point, releasing large amount of X-ray radiation. Such an accretion process is known as Roche Lobe overflow, and in Low-Mass X-ray Binaries (LMXBs), accretion occurs primarily through this mechanism (Paczyński, 1971). Stars with



(a) Roche lobe field lines in a binary system



(b) As star fills the Roche lobe, matter transfers and an accretion disk starts forming.

Figure 1.2: Schematic diagram illustrating the stages of Roche Lob overflow. [Credit: Swinburne University of Technology]

high mass shed their matter through stellar winds. For the case of High-Mass X-ray binaries (HMXBs), stellar winds ejected by the massive companion are gravitationally captured by the compact object, fueling the accretion process. The massive stars are highly luminous and so dominate the optical emission, whereas the X-ray counterpart is dominated by the compact object due to the energy released by accretion processes. Massive stars in these systems can transfer mass through Roche lobe overflow when the system is compact.

In this thesis, we primarily study low-mass X-ray Binary systems in which Roche lobe overflow is the primary accretion mechanism. The following section discusses the classification of the low-mass X-ray binary system.

1.3 Classification of Low-Mass X-ray binary

Based on the type of compact object present in the binary system, the low-mass X-ray binaries can be further classified into two categories. Given that the accretion mechanism is the same in both systems, the observational properties differ because of the physical differences between neutron stars and black holes.

- **Black Hole Low-Mass X-ray Binaries (BH LMXBs)**

  As the compact object is a black hole, there is no solid surface. Consequently, in these systems, no surface-related phenomena (such as thermonuclear X-ray bursts) are observed. Their emission is primarily governed by the accretion disk and the surrounding hot plasma (corona). These systems, when observed, are found to exhibit distinct spectral states as the source evolves in its emission properties (Remillard & McClintock, 2006).

- **Neutron Star Low-Mass X-ray Binaries (NS LMXBs)**

  In neutron star LMXBs, the presence of a solid surface leads to additional observational features, such as Type-I X-ray bursts, which result from unstable nuclear burning and pulsations due to the presence of a strong magnetic field. Based on the spectral and timing properties, neutron star low-mass X-ray binaries can be further sub-classified into distinct classes (Hasinger & Klis, 1989)

In this work, we focus on neutron star LMXBs and investigate their variability properties and classification in detail.

1.4 Outbursts in Neutron Star Low-Mass X-ray Binaries

Kiel & Hurley (2006) predicted that our Galaxy hosts more than 10^4 accreting Neutron Star X-ray Binaries. Despite this, a fraction of those have been detected so far, as these sources are not actively emitting in their quiescent (quiet) state. Such sources can only be detected when there is an outburst where the source's luminosity increases by a factor of 10^3 - 10^7 when compared with their quiescent luminosity (King, 1998). Usually, an outburst phase for a source spans a few days to months or even years, depending on the source. Many models have been proposed over time to explain the cycle of outbursts and quiescent states. Among these, the disk instability model (DIM; Meyer & Meyer-Hofmeister, 1981) provides a widely accepted framework in explaining the basic outline of an outburst. The model is based on thermal–viscous instabilities in the accretion disk, which trigger transitions between quiescent and outburst states (Done et al., 2007).

When the mass accretion rate is less than the critical accretion rate, the temperature and viscosity are initially low, and so matter starts to pile up. As a result, the disc temperature increases, and above a certain threshold, the disc material starts ionizing, increasing opacity and eventually trapping energy that heats the disc. The resulting thermal instability leads to a rapid increase in viscosity, which in turn enhances the mass accretion rate onto the compact object. This sudden increase in viscosity increases the mass accretion rate, leading to a significant release of gravitational potential energy and a corresponding increase in temperature, which is observed as an outburst. Once most of the accretion is complete, the temperature decreases, and the quiescent phase resumes (Lasota, 2001).

1.4.1 Evolution of a source during an outburst

X-ray binaries, when in outburst, are considered among the brightest objects in the sky. During this bright phase, a hot plasma region known as 'corona' forms near the compact object from the ionized accreted material. The accretion disk provides soft thermal photons to the corona, which are then Compton upscattered by the energetic, relativistic elec-

trons in the corona to higher energies (Shakura & Sunyaev, 1973; Shapiro et al., 1976; Sunyaev & Titarchuk, 1980). These upscattered photons constitute the hard X-ray component of the observed spectrum.

An outburst typically consists of two stages: the rising phase, where the accretion rate increases and so the luminosity, and the decay phase, where there is a gradual decrease in the accretion rate. However, the evolution is not described by luminosity changes alone, as sources also undergo distinct spectral and timing state transitions. At the beginning of an outburst, the source is usually in a hard state, with the emission dominated by the Comptonized component, and strong variability is observed. As the accretion rate increases, the inner edge of the accretion disk is assumed to move inward, leading to stronger thermal emission from the disk. This marks a transition to the soft state, where the disk dominates, and variability is significantly reduced. These changes in spectral and timing properties are reflected in observable quantities that can be used to track the source's evolution during an outburst.

1.4.2 Classification of Neutron Star Low-Mass X-ray Binaries

For the Low-mass X-ray binaries, the Color-Color Diagram (CCD) (See Section 3.2.1) is often used to track the spectral evolution (Hasinger & Klis, 1989). They give the variation of hard X-ray hardness ratio (ratio of count rates in "high" energy band to "mid" energy band) to the soft X-ray hardness ratio (ratio of count rates in "mid" energy band to "low" energy band) (for the analysis done in this work, the specific values adopted for the low, mid, and high energy bands are provided in Section 3.2.1).

Furthermore, Neutron Star Low-Mass X-ray binary systems (NS LMXBs) can be subdivided into two classes based on the shape of the track traced in their color-color diagrams: Atoll sources and Type-Z sources. In the following subsections, characteristics of each subclass are described separately.

Atoll Sources

Atoll sources are a class of NS LMXBs that show characteristic tracks in the color–color and hardness–intensity diagrams, typically tracing a "C"-shaped pattern in the color–color diagram. Based on their position along this track, they are classified into different spectral states.

In the color–color diagram (CCD), Atoll sources occupy the Island

State (IS), Extreme Island State (EIS), and the Banana branch. The Banana state is further divided into the Upper Banana (UB), Lower Banana (LB), and Lower Left Banana (LLB).

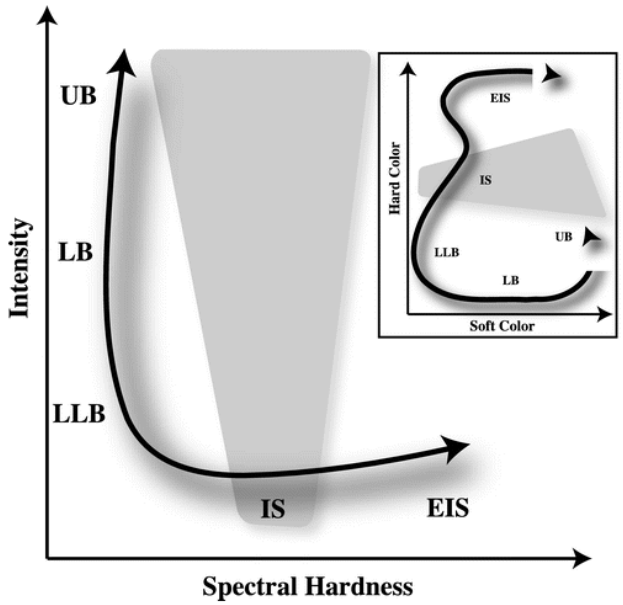


Figure 1.3: HID and CCD representation of a typical Atoll type LMXB. Image reference: Klein-Wolt & Van Der Klis (2008)

Fig. 1.3 gives a schematic diagram of the Hardness-intensity plot(see sec. 3.5.1) (plot of hard X-ray hardness ratio vs intensity) and CCD for Atoll Sources with the branch labeling. Generally, as the source moves from the extreme island branch to the upper banana branch through the island, lower left banana, and lower banana, the spectral state softens and the intensity increases.

Z-type Source

Sources that trace out a "Z-shaped" path in the CCD are referred to as "Type-Z" sources. The Z track splits into the Horizontal branch, the Normal branch, and the Flaring branch. Fig 1.4 provides a sample CCD plot, labeled with the branches. The Type-Z sources are highly variable in nature, and they tend to cover the entire Z track in periods ranging from hours to a few days (Hasinger et al., 1990). It is expected that the luminosity and the accretion rate to increase as the source moves from the horizontal branch to the flaring branch (Migliari & Fender, 2006).

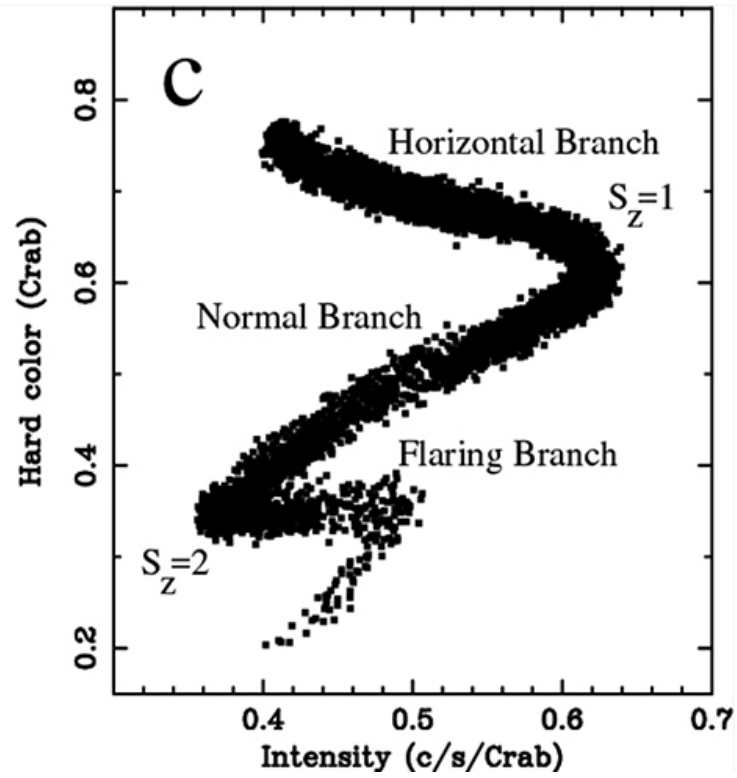


Figure 1.4: CCD plot of Type-Z source using RXTE data. Image reference: Chaty (2022)

Accreting Millisecond X-ray Pulsars

Sources with neutron stars having strong magnetic fields ($\sim 10^8$ G) (Mukherjee et al., 2015) can channel the inflowing matter along the magnetic field lines at the magnetic poles. As a result, a regular repeated signal is produced with the X-ray emission and is observed as pulsations. These sources are referred to as Accreting Millisecond X-ray Pulsars (hereafter referred to as Pulsars for classification purposes).

Atoll + Type-Z sources

There are sources that exhibit mixed behavior between Atoll and Type-Z on the CCD/HID track. Homan et al. (2010a) showed that in sources like *XTE J1701-462*, the spectral and timing properties evolve from Atoll-like behavior at lower intensities to Type-Z behavior at higher intensities. This indicates that the source does not belong strictly to a single subclass, but instead shows characteristics of both. Such sources can therefore be classified as Atoll + Type-Z.

1.5 Motivation

The NS LMXBs are highly variable and transient systems. Their transient nature implies that they spend long durations in a quiescent state, followed by episodic outbursts during which the luminosity increases by several orders of magnitude (Campana et al., 1998; Yan & Yu, 2015). These outbursts are accompanied by strong variability and significant changes in spectral behavior. Such type of variability has been observed in several transient neutron star systems such as *Aql X-1*, *SAX J1808.4-3658* and *4U 1608-52*, where repeated outbursts are observed where they showcase dramatic transitions in the spectral behavior (Güngör et al. (2014); Markwardt et al. (2002) & Güngör et al. (2014)). Importantly, even for the same source, it is observed for some cases that the spectral evolution is not reproducible across outbursts. *Aql X-1* has been observed to trace different hardness–intensity and color–color tracks in different outbursts, indicating that the accretion geometry and emission properties can vary significantly from one event to another (Güver et al., 2022; Maitra & Bailyn, 2004).

The usual outburst duration varies from several weeks to months or even a year, depending on the source. In addition to such short-term variability, some systems also exhibit unusual long-term behavior; for instance, *EXO 0748–676* remained in an active outburst state for more than two decades before returning to quiescence (Degenaar et al., 2014). Bhattacharya & Bhattacharyya (2025) found that the standard thermal-viscous instability framework cannot account for these types of long-duration outbursts, indicating the need for additional mechanisms to explain their origin.

The emissions observed during the outburst phase also provide valuable insights for the disk-corona geometry of the system. Moreover, spectral state transitions from Hard to Soft are often linked to the appearance of radio jets, suggesting a strong correlation between the inner disk, corona, and the jet-producing regions (Migliari et al., 2007).

The distinct tracks traced by Atolls and Type-Z sources in their hardness-intensity diagrams and color-color diagrams are not only observational features, but also a strong representation of the physical processes occurring near the compact object. However, this classification is not always rigid. Sources such as *XTE J1701–462* have been observed to

exhibit both Type-Z–like and Atoll-like behavior within a single outburst, suggesting that these categories may represent highly variable accretion phenomena (Homan et al., 2010a). Furthermore, even within the Atoll class, multiple transition paths have been observed in hardness-intensity and color-color diagrams during state evolution (Gladstone et al., 2007a), suggesting that the mapping between spectral states and physical conditions is complex and non-unique. This illustrates the complexity and time variability of these systems.

Altogether, these observational features highlight that NS LMXBs are not only highly variable and transient but also exhibit complex, non-uniform behavior across both short and long timescales. Interestingly, despite exhibiting such volatile behavior, when the X-ray data from multiple sources are represented in a Color-Color-Intensity (CCI) plot (a three-dimensional plot with hard color, soft color, and intensity on three axes), clustering is observed (Vrtilek & Boroson, 2013).

Thus, in light of the above discussion, a data-driven approach that can capture this intrinsic complexity can be employed. In particular, machine learning techniques can be used to identify the source class and spectral state (i.e., branch in the CCD track) for a given observation, thereby providing a deeper understanding of these variable systems. In this work, we utilize machine learning methods based on spectral and timing features extracted from X-ray data to classify NS LMXBs and extend this approach to cluster their states within a given source class. This dual-level classification helps provide a comprehensive understanding of that particular observation.

Importantly, this approach facilitates understanding and predicting the evolution of newly detected outbursts and can also be helpful in timely multi-wavelength follow-up observations.

Hence, it is essential to review the existing approaches to classifying the source class and spectral behavior of LMXBs. In the following section, an overview of studies in this field is presented, highlighting the progress achieved to date and setting the stage for the present work.

1.5.1 Review of related works

Vrtilek & Boroson (2013) suggested a model that plotted ellipsoids around the centroid of each cluster in the CCI diagram and then minimized their volume to obtain clustering based on the type of compact

object present in the binary system. This was one of the first analyses done in this field and provided a model-independent approach to classification. This model effectively distinguished between classes with jets and those without; however, there was significant overlap in some regions of the plot, especially between the Black hole and the Neutron star, resulting in inefficient clustering. To overcome this limitation, Gopalan et al. (2015) applied a Bayesian probabilistic model to the CCI data to improve classification. This method provides classification probabilities and associated uncertainties by using a Gaussian process to smooth the data while preserving the underlying geometric structure. It was observed that the model may misclassify boundary regions and is computationally expensive, as the Gaussian process's computational cost scales as $O(N^3)$, where N is the number of input data points.

Pattnaik et al. (2020) performed an analysis to classify LMXBs based on the nature of the compact object using machine learning (ML) algorithms. For each observation, they used 43 count-rate values across the 5–25 keV energy range as the input vector for the classification algorithms. The methods used included Random Forest, Decision Tree, Logistic Regression, XGBoost (boosting), K-Nearest Neighbors, and Support Vector Machine classifiers. Cross-validation results indicated that Random Forest provided the best performance, with $87 \pm 13\%$ accuracy in distinguishing between BH LMXBs and NS LMXBs. It should be noted that all the studies mentioned above used observational data provided from the *RXTE* mission. Sreehari & Nandi (2021) employed ML methods for the first time on multi-mission observations for the classification of accretion states of BH-X-ray binaries. Archival data from *RXTE*, *Swift*, *MAXI*, and *AstroSat* were used to extract features like X-ray flux, hardness ratios, presence of QPOs, photon indices, and disc temperature. They applied clustering algorithms, such as K-means clustering and Hierarchical clustering, to classify the accretion state of each outburst.

There are a few studies that use ML methods to explore the long-term variability and evolution of the BH X-ray binary, GRS 1915+105. Ricketts et al. (2023) used an unsupervised learning approach, using an autoencoder to classify the variability patterns in the light curve of this source. Huppenkothen et al. (2016) performed supervised learning on the classes of the variability patterns in the light curve, using Logistic

Regression with an 85% accuracy score and Random Forest with a 91% accuracy score.

It is therefore clear that ML methods have proven efficient in classifying both the source class and the spectral state of the X-ray binaries. However, most previous work focused on missions such as RXTE, primarily classifying sources based on the type of compact object in the binary system. Thus, NS LMXBs have remained relatively unexplored using this approach, particularly with *AstroSat* data. This presents a strong opportunity to extend ML-based classification methods to these systems, to explore their state transitions, and to examine clustering behavior across different source classes.

1.5.2 Objectives

In the present work, we use archival data from the *AstroSat* mission for 35 NS LMXBs. We applied both the supervised and unsupervised approaches to get classification in two stages:

- The source class into four categories: Atoll, Type-Z, AMXP, and Atoll + Z (showing mixed behavior of both Atoll and Type-Z sources).
- Spectral state classification within a source class.

The specific objectives of this study are outlined as follows :

1. **Data Processing**: Process *AstroSat* observations of 35 NS LMXB sources and assign source class labels to each source, based on the literature.
2. **Visualization**: Construct 3D CCI plots and Power-Color diagrams to confirm the clustering behavior among different source classes visually.
3. **Source class Classification**: Apply suitable supervised and unsupervised ML algorithms, tune hyperparameters, and classify source classes using extracted features.
4. **State Identification and Interpretation**: Incorporate spectral state information into the feature space to perform clustering, identify state evolution, and obtain physical inference from the results to understand the underlying behavior of the systems.

Chapter 2

Instrument

AstroSat is India's first dedicated multi-wavelength astronomy mission, designed to study high-energy processes, detect X-ray sources, and perform deep ultraviolet observations. The satellite has a launch mass of 1,513 kg and was placed in a 650 km orbit with an inclination of 6^o on 28 September 2015 from the First Launch Pad at the Satish Dhawan Space Centre, operated by ISRO. It has a total of 14 orbits per day with an orbital period of 97.6 minutes.

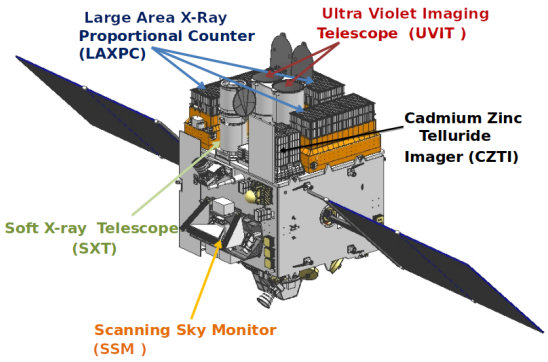


Figure 2.1: AstroSat satellite carrying five payloads
[Image reference: https://astrobrowse.issdc.gov.in/astro_archive/archive/astrosat_images/astrosat_wireframe.png]

AstroSat spacecraft has the following five payloads:

- Ultra Violet Imaging Telescope (UVIT): Consists of two 380 mm telescopes. One covers the Far Ultra Violet band (130–180 nm) and the other covers the Near Ultra Violet (200–300 nm) and visible (320–550 nm) bands.
- Large Area X-ray Proportional Counter (LAXPC): Provides a large effective area (~8000 cm^2) suitable for fast X-ray timing studies.

- Cadmium–Zinc–Telluride Imager (CZTI): Operates in the hard X-ray range (20–100 keV) with coded-mask imaging and also supports X-ray polarimetry.

- Soft X-ray Telescope (SXT): Covers the 0.3–8 keV energy range using focusing mirrors and an X-ray CCD for soft X-ray imaging and spectroscopy.

- Scanning Sky Monitor (SSM): An all-sky monitor operating in the 2.5–10 keV band for detecting and tracking transient X-ray sources.

For the work presented in this thesis, we used archival data from the LAXPC instrument onboard the *AstroSat* satellite[1].

2.1 LAXPC instrument

The Large Area X-ray Proportional Counter (LAXPC) is one of the major payloads on *AstroSat*, covering the 3–80 keV energy range.
The LAXPC instrument is expected to provide a wealth of data on a variety of astrophysical objects in this band, with a fine timing resolution of 10 μs.



Figure 2.2: Schematic showing the three identical LAXPC units onboard *AstroSat*.

[Image reference: Yadav et al. (2016)]

The three LAXPC units onboard *AstroSat* together provide a large effective area and high sensitivity. The total effective area of the three detectors is about 6000 cm^2 in the 5–20 keV range.

[1] https://astrobrowse.issdc.gov.in/astro_archive/archive/Home.jsp

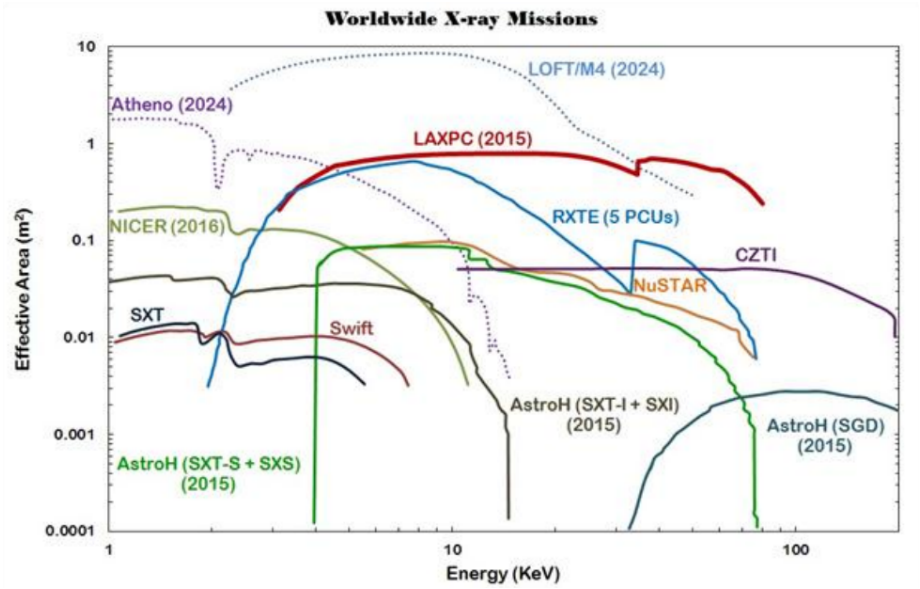


Figure 2.3: Comparison of the effective area of LAXPC and the other instruments of ASTROSAT with all the other dedicated astronomical X-ray missions [Image reference:https://www.issdc.gov.in/docs/as1/AstroSat_Payloads.pdf]

Each detector is filled with xenon gas along with a small amount of methane, and has a large detection volume (15 cm depth). This results in a detection efficiency greater than 50% in the 30–80 keV band. LAXPC is particularly notable for its excellent timing resolution, enabling studies of extremely fast variability in X-ray sources. Its energy resolution is approximately 12% FWHM at 30 keV, making it one of the most sensitive instruments for X-ray astronomy. m

Chapter 3

Observations and Data Preparations

The primary aim of this thesis is to classify Neutron Star LMXB sources based on their spectral and timing behavior, and further perform a finer classification of spectral states for a particular source class (Atolls). To maintain uniformity in the features among different classes and observations, a systematic sample selection procedure was adopted.

In this chapter, we discuss in detail the data processing, sample selection, feature extraction, and feature space visualization.

3.1 Data Processing

3.1.1 Sample Selection

We searched the entire ISRO Science Data Archive for the AstroSat mission[1] for bursting and non-bursting Neutron Star LMXB systems. In total, 35 distinct NS-LMXBs were sampled, covering the major source classes, including Atoll, Pulsar, Type-Z, and hybrid Atoll + Type-Z sources. This provides a well-sampled starting point for our analysis.

The observational data used in this analysis were obtained from the LAXPC instrument onboard AstroSat. The instrument consists of three similar Proportional Counter Units - LAXPC10, LAXPC20, and LAXPC30, with a working energy range of 3 - 80 keV. It has been reported that LAXPC10 and LAXPC30 experienced gain instability arising due to gas leakage, from March 2018 onward. Therefore, for the present analysis, data from only LAXPC20 were considered.

[1] astrobrowse.issdc.gov.in/astro_archive/archive/Home.jsp

The level-1 data from the AstroSat Archive was processed using the LAXPCSoftware[2] (15 Aug, 2022 version). Using the standard procedure of LAXPC analysis (available in LAXPCSoftware[3]), respective GTI files, spectrum, and background spectrum were generated. The data is collected in Event Analysis (EA) mode, which includes time, channel number, and anode ID for each event in the event file.

Table 3.1 gives an Observation log of the sources included in the sample.

Table 3.1: Observation details of selected X-ray binary sources

No	Source Name	Total Obs.	Observation ID	Observation Date	Exp. Time (ks)
1	4U 0614+09	2	A02_175T01_9000000874	10-12-2016	10.8
			A02_175T01_9000000976	17-01-2017	27
2	4U 1246-588	1	G07_065T02_9000001150	08-04-2017	44.8
3	4U 1323-62	1	A02_134T01_9000001036	16-02-2017	49.5
4	4U 1254-690	1	A04_162T01_9000002074	03-05-2018	99.7
5	4U 1543-624	1	G06_032T01_9000001074	07-03-2017	60
6	4U 1608-52	2	T03_214T01_9000003758	14-07-2020	51
			G05_140T01_9000000628	28-08-2016	40
7	4U 1626-67	5	G07_049T01_9000001352	02-07-2017	67.85
			G02_020T02_9000000268	13-01-2016	8.21
			G08_084T01_9000002100	15-05-2018	36.35
			G02_020T02_9000000294	26-01-2016	12.19
			G05_021T01_9000000624	26-08-2016	41.69
8	4U 1636-536	11	G05_195T01_9000000530	02-07-2016	35.9
			A04_055T01_9000001574	02-10-2017	48.23
			G08_033T01_9000002278	06-08-2018	10.82
			G08_033T01_9000002084	09-05-2018	10.82
			G05_195T01_9000000598	13-08-2016	36
			P01_175T03_9000000322	15-02-2016	38.42
			T02_086T01_9000002316	18-08-2018	104
			G08_033T01_9000001906	19-02-2018	13.25
			G05_002T01_9000000616	20-08-2016	33.07
			G07_040T01_9000001326	21-06-2017	19.52
			G06_104T01_9000001060	28-02-2017	42.42
9	4U 1702-429	2	A06_002T03_9000003080	08-08-2019	57
			A04_225T01_9000002062	27-04-2018	8.19

Continued on next page

[2] http://astrosat-ssc.iucaa.in/laxpcData

[3] http://astrosat-ssc.iucaa.in/uploads/threadsPageNew_SXT.html

Table 3.1 continued

No	Source Name	Total Obs.	Observation ID	Observation Date	Exp. Time (ks)
10	4U 1705-44	1	A09_044T01_9000003888	18-09-2020	10
11	4U 1728-34	7	G08_035T01_9000002268	01-08-2018	23.88
			A05_229T02_9000002890	04-05-2019	206
			G05_190T01_9000000578	05-08-2016	64.7
			T01_041T01_9000000362	07-03-2016	29.8
			G08_035T01_9000002234	17-07-2018	21.07
			G08_035T01_9000001904	19-02-2018	21.22
			A05_229T02_9000003060	25-07-2019	39.65
12	4U 1735-44	6	A05_062T01_9000002892	06-05-2019	19
			A10_059T01_9000004634	12-08-2021	12.23
			A03_039T01_9000001462	13-08-2017	37.47
			A10_059T01_9000004656	24-08-2021	9.69
			A05_062T01_9000003132	28-08-2019	18.7
			A04_126T01_9000002262	31-07-2018	46.8
13	4U 1812-12	1	A04_223T01_9000002258	26-07-2018	50.0
14	4U 1820-30	10	A04_055T02_9000001576	03-10-2017	13.36
			G08_036T01_9000001584	05-10-2017	8.52
			G08_036T01_9000002290	09-08-2018	10.48
			A03_107T01_9000001530	09-09-2017	1.48
			G08_036T01_9000002034	11-04-2018	9.1
			T05_022T01_9000005070	15-04-2022	44.36
			A02_098T01_9000000736	17-10-2016	67.3
			A03_107T01_9000001246	26-05-2017	12.69
			A03_107T01_9000001418	30-07-2017	15.7
			G07_047T01_9000001424	31-07-2017	22
15	4U 1850-087	1	G07_065T01_9000001454	10-08-2017	35.6
16	Cir X-1	3	G08_043T01_9000002288	09-08-2018	17.36
			G08_043T01_9000001902	18-02-2018	11
			G06_083T01_9000000984	26-01-2017	41.25
17	Cyg X-2	10	A06_002T02_9000002982	09-06-2019	71.54
			A13_037T01_9000006432	16-09-2024	16.4
			A13_037T01_9000006256	17-05-2024	23.4
			A13_037T01_9000006300	18-06-2024	24.25
			T01_058T01_9000000348	28-02-2016	62.8
			A13_037T01_9000006318	28-06-2024	8
			A06_002T02_9000003206	28-09-2019	84.67
			G08_025T01_9000002130	29-05-2018	95.5
			A06_002T02_9000003064	29-07-2019	14.5
			A13_037T01_9000006212	30-04-2024	9.5

Continued on next page

Table 3.1 continued

No	Source Name	Total Obs.	Observation ID	Observation Date	Exp. Time (ks)
18	EXO 0748-676	1	T05_213T01_9000006344	24-07-2024	30
19	GS 1826-238	2	T01_042T01_9000000366	09-03-2016	90.63
			G05_002T03_9000000608	17-08-2016	5.11
20	GX 13+1	1	G06_032T02_9000000730	11-10-2016	79.06
21	GX 17+2	10	G08_037T01_9000002264	01-08-2018	20
			A05_062T02_9000003138	01-09-2019	20
			G08_037T01_9000001588	06-10-2017	18
			A09_044T02_9000003814	08-09-2020	7
			T02_087T01_9000002352	10-09-2018	50
			G05_112T01_9000000452	11-05-2016	99
			A09_044T02_9000003814	14-08-2020	7
			A03_072T01_9000001484	22-08-2017	40
			G08_037T01_9000002256	26-07-2018	20
			G08_037T01_9000001928	28-02-2018	18
22	GX 3+1	4	G08_039T01_9000001582	05-10-2017	9.57
			G08_039T01_9000002292	09-08-2018	7.77
			G08_039T01_9000002036	12-04-2018	8.05
			A04_122T01_9000002064	29-04-2018	41.15
23	GX 340+0	13	A06_009T02_9000002882	01-05-2019	5.3
			G08_022T01_9000002078	06-05-2018	68.07
			A06_009T02_9000003076	06-08-2019	3.33
			A06_009T02_9000003152	10-09-2019	3.89
			A10_059T02_9000004636	12-08-2021	41.4
			T05_218T01_9000006386	13-08-2024	15
			A06_009T02_9000003170	17-09-2019	4.8
			A09_134T01_9000003896	19-09-2020	99.27
			T05_188T01_9000006138	23-03-2024	23.01
			A06_009T02_9000003194	23-09-2019	7.4
			A06_009T02_9000003130	28-08-2019	4.6
			A06_009T02_9000003006	29-06-2019	4.5
			G07_016T01_9000001420	30-07-2017	38.31
24	GX 349+2	7	A13_037T05_9000006216	02-05-2024	64.5
			A13_037T05_9000006332	10-07-2024	41.16
			A06_009T05_9000003154	10-09-2019	33.14
			A02_071T01_9000001078	12-03-2017	162.53
			A03_029T01_9000001384	14-07-2017	111.67
			A06_009T05_9000003164	16-09-2019	7
			A06_009T05_9000003196	23-09-2019	23
25	GX 5-1	9	A06_009T03_9000003072	05-08-2019	4.3
			G08_068T01_9000001580	05-10-2017	9.8

Continued on next page

Table 3.1 continued

No	Source Name	Total Obs.	Observation ID	Observation Date	Exp. Time (ks)
			G08_068T01_9000002090	10-05-2018	4.2
			A06_009T03_9000003098	17-08-2019	099
			A06_009T03_9000003198	24-09-2019	5.01
			A06_009T03_9000003122	25-08-2019	18.14
			G06_114T01_9000001056	26-02-2017	20.5
			G08_068T01_9000001922	26-02-2018	20.5
			A06_009T03_9000003128	28-08-2019	5.2
26	IGR J00291+5934	1	A07_006T04_9000003210	20-10-2019	30.1
27	IGR J17062-6143	1	A03_050T01_9000001468	16-08-2017	49.99
28	IGR J17591-2342	2	T02_088T01_9000002320	23-08-2018	34.7
			T02_088T01_9000002332	27-08-2018	37.7
29	MAXI J0911-655	3	T04_025T01_9000004538	14-07-2021	50
			A05_206T01_9000002518	15-11-2018	50.9
			A05_206T01_9000002528	20-11-2018	47.07
30	MXB 1658-298	3	A02_158T01_9000001064	01-03-2017	21.26
			A02_158T01_9000001080	14-03-2017	3.5
			A02_158T01_9000001184	18-04-2017	18.26
31	UW crb	1	A03_074T01_9000001192	20-04-2017	72.6
32	XTE J1701-407	2	A03_095T03_9000001474	19-08-2017	13.53
			T01_157T01_9000001050	24-02-2017	5.91
33	XTEJ1701-462	2	T05_059T01_9000005354	01-10-2022	85
			T05_054T01_9000005350	25-09-2022	50
34	XTE J1710-281	4	A03_084T01_9000001440	07-08-2017	3.09
			A03_084T01_9000001382	14-07-2017	12.5
			A03_084T01_9000001188	19-04-2017	11
			A03_084T01_9000001544	21-09-2017	11
35	XTE J1739-285	2	T03_143T01_9000003208	01-10-2019	94.5
			T03_180T01_9000003524	19-02-2020	88.5

3.1.2 Data Reduction

The light curves were generated from the event file for each observation using a time bin size of 1s. The data were cleaned by using `sigma_clip` function of *Astropy* library, by applying a 3σ confidence threshold. The sub-sampling was repeated for 5 iterations to make sure that no outliers are left.

The generated light curves were **background subtracted** using the

`heasoft` tool `lcmath`[4]. In addition, the data processing pipeline ensures that the light curves are corrected for instrumental dead time.

In total, the final dataset used for source-class-based classification consists of **35 sources** and **133 observations**.

3.2 Feature Extraction

In this section, we discuss in detail the features and procedures used to extract them.

3.2.1 Hardness Ratios

The Hardness Ratio (Hasinger & Klis, 1989) is defined as the ratio of high energy flux to the low energy flux. For our analysis, we compute two hardness ratios, Hard color and Soft color. The definitions of Hard color and Soft color are given in Eq. 3.1.

$$\begin{aligned} \text{Hard Color} &= \frac{\text{Count Rate in High energy range}}{\text{Count Rate in Medium energy range}} \\ \text{Soft Color} &= \frac{\text{Count Rate in Medium energy range}}{\text{Count Rate in Low energy range}} \end{aligned} \tag{3.1}$$

Park et al. (2006) gave another formula for calculating the hardness ratios, which is given in Eq. 3.2. They also found that the ratio of the fractional differences in count rates works well in both high and low count regimes when multiple observations from different sources are plotted on a single plot.

$$\begin{aligned} \text{Hard Color} &= \frac{\text{High - Low}}{\text{High + Low}} \\ \text{Soft Color} &= \frac{\text{Medium - Low}}{\text{Medium + Low}} \end{aligned} \tag{3.2}$$

where High, Medium, and Low mean the Count rates in the high energy range, the medium energy range, and the low energy range, respectively. The values 4-6 keV, 6-10 keV, and 10-20 keV are selected for the low, medium, and high energy ranges, respectively, for our analysis.

- Please note that in this analysis, the **spectral hardness parameters** are obtained using the **energy-resolved background-subtracted light curves**.

[4] https://heasarc.gsfc.nasa.gov/docs/software/xronos/manual/node6.html

For this analysis, both definitions were initially considered. However, later, we adopted the traditional definition given in Eq. 3.1. For further analysis, the other definition (Eq. 3.2) exhibited complex behavior and patterns that were not suitable for comparison with existing nomenclature.

The software used individual response files for each observation and obtained calibration for channel number and energy range. From this, the count rates were calculated by taking the ratio of photon counts in the corresponding energy range to the bin size.

The Hardness ratio values were found to be sensitive to the chosen bin size, as observed during the error analysis. If the bin size is not optimized, the propagated error in the Hardness ratio increases significantly. Since the analysis involves multiple sources, it is important to set the bin size to provide more reliable estimates of feature values with reduced error. To illustrate this, sources with extreme behavior (in outbursts) were investigated to show how the error in the Hardness Ratio varies with bin size.

The analysis suggested that beyond the bin size of 512 s, the (relative) error is low and stable. Thus, 512 s was chosen as the optimized bin size for subsequent analysis. The Hard and Soft colors for the 512 s bin size were then calculated for all sources, using the defined energy ranges and definitions as per the Eq. 3.1. These computed values constitute two key features incorporated into our dataset.

3.2.2 Intensity

The intensity values were computed by summing the count rates over the entire energy range, i.e., 4-20 keV. Moreover, to achieve consistent comparisons of observations across different sources at various distances, it is necessary to apply appropriate normalization to the count rates. This normalization was performed using the distance-luminosity relation, assuming that the observed intensity (count rates) is inversely proportional to the source's distance.

The normalized count rate I_{norm} was then calculated using Eq. (3.3).

$$I_{norm} = I_{obs}\left(\frac{d}{d_{ref}}\right)^2 \tag{3.3}$$

where I_{obs} is the observed count rate, d is the distance to the source, and d_{ref} = 5 kpc is the distance adopted as a reference for normaliza-

tion. The distances to each source are taken from the literature, and the normalization constant is calculated. Table 3.2 lists these normalization constants along with distance information and references.

Table 3.2: List of selected neutron star low-mass X-ray binary sources along with their distances and normalization factor $(d/5)^2$.

Sr. No	Source Name	Distance (kpc)	Normalization Constant	Ref.
1	4U 0614+09	3.2 ± 0.2	0.41	(1)
2	4U 1246-588	4.3 ± 0.6	0.74	(2)
3	4U 1254-69	13 ± 3	6.76	(3),(4)
4	4U 1323-62	10 – 20	9.00	(5),(6)
5	4U 1543-624	~ 7	1.96	(7),(8)
6	4U 1608-52	2.9 – 4.5	0.55	(9),(12)
7	4U 1636-53	6.0 ± 0.5	1.44	(10)
8	4U 1702-429	~ 5.6	1.25	(11),(12)
9	4U 1705-44	7.4	2.19	(12)
10	4U 1728-34	4.4 – 5.1	0.90	(13)
11	4U 1735-44	6.5 ± 1.0	1.69	(33)
12	4U 1812-12	2.3 – 4.6	0.48	(14),(15)
13	4U 1820-30	8.0 ± 0.3	2.56	(16)
14	4U 1850-087	8	2.56	(17)
15	EXO 0748-676	5.9 – 7.7	1.85	(18)
16	GS 1826-24	6.07 ± 0.18	1.47	(19)
17	MXB 1658-298	10	4.00	(20)
18	UW CrB	11 – 24	12.25	(21)
19	XTE J1701-407	5.0 ± 0.4	1.00	(31)
20	XTE J1710-281	12 ± 1.8	5.76	(12)
21	XTE J1739-285	4	0.64	(32)
22	GX 3+1	$10.17^{+0.07}_{-0.06}$	4.14	(34)
23	GX 5-1	4.2	0.71	(35)
24	GX 17+2	8.5 ± 1.2	2.89	(22)
25	GX 349+2	9.2	3.39	(23)
26	GX 340+0	11.0 ± 3.3	4.84	(38)
27	Cyg X-2	7.2 ± 1.1	2.07	(36)
28	IGR J00291+5934	3.1	0.38	(24)
29	IGR J17062-6143	7.3 ± 0.5	2.13	(25)
30	IGR J17591-2342	7.6	2.31	(26)
31	MAXI J0911-655	9.5	3.61	(27)
32	4U 1626-67	10	4.00	(37)
33	Cir X-1	9.4	3.53	(28),(30)
34	GX 13+1	7 ± 1	1.96	(22)
35	XTE J1701-462	8.8 ± 1.3	3.10	(29)

References:

(1) Chen et al. (2022); (2) Zand et al. (2008); (3) Bhattacharyya (2007); (4) In 'T Zand et al. (2003); (5) Balman (2010); (6) Parmar et al. (1989); (7) Ludlam et al. (2019); (8) Wang & Chakrabarty (2004); (9) Bhattacherjee et al. (2024a); (10) Galloway et al. (2006); (11) Iaria et al. (2016); (12) Galloway et al. (2008); (13) Kashyap et al. (2025); (14) Armas Padilla et al. (2022); (15) Cocchi et al. (2000); (16) Ballan-

tyne & Strohmayer (2004); (17) Lu et al. (2024); (18) Hynes & Jones (2009); (19) Heger et al. (2007); (20) Sharma et al. (2018); (21) Jonker et al. (2003); (22) Thomas et al. (2025); (23) Grimm et al. (2002); (24) De Falco et al. (2017); (25) Keek et al. (2017); (26) Sanna et al. (2018); (27) Gasealahwe et al. (2023); (28) Heinz et al. (2013); (29) Rankin et al. (2024); (30) Gasealahwe et al. (2023); (31) Falanga et al. (2009); (32) Bailer-Jones et al. (2018); (33)Lavanya et al. (2024a); (34)Thomas et al. (2023a); (35) Clark (2018); (36) Orosz & Kuulkers (1999); (37) Juris et al. (2025); (38) Fender & Hendry (2000)

Thus, the normalized intensity values were subsequently added to the sample space along with Hard color and Soft color as features. Fig. 3.1 provides the number of data points per source used in the analysis.

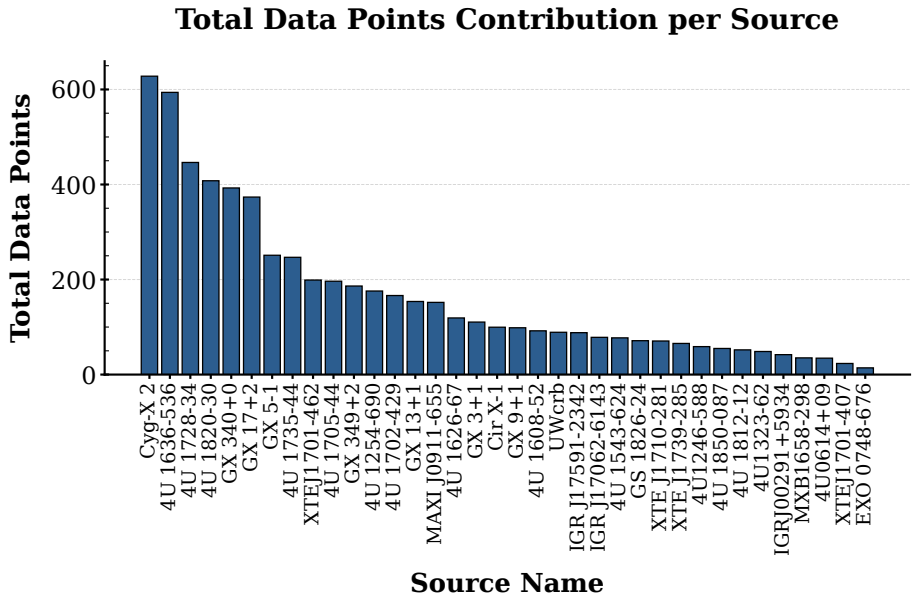


Figure 3.1: Number of data points per source used in the analysis. Each data point corresponds to a 512 s segment of the light curve, from which the hard color, soft color, and intensity are extracted.

3.3 Power Color and Fractional RMS Amplitude

Grinberg et al. (2014) showed that the Power Density Spectrum (PDS) and fractional RMS variability are strongly dependent on the spectral state of an X-ray binary source. The PDS provides information about the variability of sources across different time scales (frequencies). The fractional RMS variability estimates the magnitude of variability of X-ray light curve relative to its average flux. In the frequency domain, it is

calculated as the square root of integrated power of the PDS curve (area under the PDS curve), after subtracting the Poisson noise contribution, over a given frequency band.

Similar to how hardness ratios were calculated in the energy space, Heil et al. (2015) introduced a method to calculate the power color features in the frequency domain. Power color features are defined as the ratios of variability amplitudes across different frequency bands. We followed the exact procedure for computing the power color values as described by Heil et al..

Thus, using the standard subroutine `laxpc_find_freqlag`, the PDS were generated using 256s integration time and 16 Hz sampling frequency for energy range 4-13 keV, with a signal-to-noise ratio threshold of 5. The generated power spectra were fitted with a model comprising multiple Lorentzians using the `XSPEC` package of `heasoft`. The area under the curve (variance) was calculated for different frequency bands, and the ratios were taken to obtain the Power Color values as defined in the following equation

$$\begin{aligned} \text{PC1} &= \frac{\text{Variance in 0.25-2.0 Hz}}{\text{Variance in 0.0039-0.031 Hz}} \\ \text{PC2} &= \frac{\text{Variance in 0.031-0.25 Hz}}{\text{Variance in 2-16 Hz}} \end{aligned} \tag{3.4}$$

By plotting the Power Color values on a logarithmic scale, an elliptical pattern is obtained, representing the color wheel, where different spectral states occupy different regions of this ellipse. This representation is known as the Power Color Diagram (PCD, see fig 3.2).

We extracted the Power Color values for each observation and the Total fractional RMS amplitude (total variance in 0.0039-16 Hz band) and added them as features in the final sample space. The final dataset consists of six features: hard color, soft color, normalized intensity, power color 1 (PC1), power color 2 (PC2), and total RMS variability amplitude.

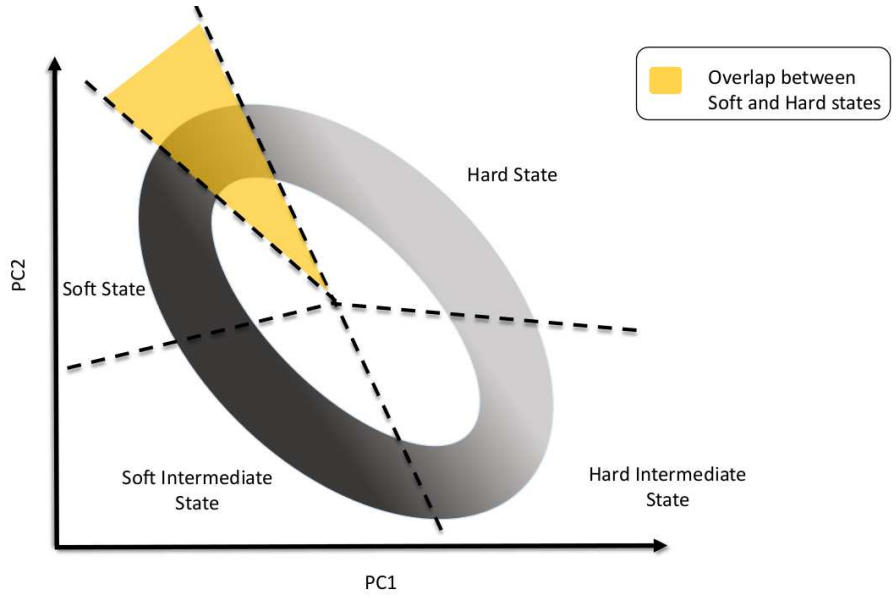


Figure 3.2: Schematic representation of spectral states in the power color diagram (PC1–PC2 plane). For the representation of the ellipse, the center point was chosen as [4.519,0.453] by the authors. [Image reference: Heil et al.]

- *It is important to note that Fourier-based techniques are required to extract power-color and RMS features, which require a sufficiently large dataset for reliable averaging. Averaging is important to ensure that the required SNR threshold is met. Thus, the power spectral features are computed over the entire observation for the present analysis and then mapped to every 512 s binned hardness ratios and normalized intensity features for the same observation.*

3.4 Dataset Labeling

3.4.1 Source-class Classification

To use supervised machine learning methods and cross-validate results from unsupervised machine learning methods, the feature sample space should be labeled.

Thus, based on the reported spectral behavior of the sources, each of them is now classified into one of the four classes, which are respectively

1. Atoll
2. Type-Z
3. Pulsar

4. Atoll + Type-Z

Please note that the Pulsar class has overlap with the atoll class, however since the physical accretion mechanism is distinctly different for them resulting in specific observable characteristics, we treat it as a separate class. The source class labeling is listed along with the source names in Table 3.3.

Table 3.3: Classification of X-ray binary sources used in this work

Sr. No.	Source Name	Source Class
1	4U 0614+09	Atoll
2	4U 1246-588	Atoll
3	4U 1254-69	Atoll
4	4U 1323-62	Atoll
5	4U 1543-624	Atoll
6	4U 1608-52	Atoll
7	4U 1636-53	Atoll
8	4U 1702-429	Atoll
9	4U 1705-44	Atoll
10	4U 1728-34	Atoll
11	4U 1735-44	Atoll
12	4U 1812-12	Atoll
13	4U 1820-30	Atoll
14	4U 1850-087	Atoll
15	EXO 0748-676	Atoll
16	GS 1826-24	Atoll
17	MXB 1658-298	Atoll
18	UW CrB	Atoll
19	XTE J1701-407	Atoll
20	XTE J1710-281	Atoll
21	XTE J1739-285	Atoll
22	GX 3+1	Atoll
23	GX 5-1	Type-Z
24	GX 17+2	Type-Z
25	GX 349+2	Type-Z
26	GX 340+0	Type-Z
27	Cyg X-2	Type-Z
28	IGR J00291+5934	Pulsar
29	IGR J17062-6143	Pulsar
30	IGR J17591-2342	Pulsar
31	MAXI J0911-655	Pulsar
32	4U 1626-67	Pulsar
33	Cir X-1	Atoll + Z
34	GX 13+1	Atoll + Z
35	XTE J1701-462	Atoll + Z

3.4.2 Spectral State based Classification

Moreover, we aim to classify the spectral state of a specific source class. Studies suggest that Type-Z sources, being highly variable, cover the entire Z track over timescales ranging from hours to a few days (Hasinger et al., 1990). In this scenario, it is challenging to obtain precise information about a single spectral state, as the tracks in the HID and CCD provide only a brief indication of the source's current state. Also, the PDS generation assumes that the light curve properties are statistically stationary. Due to the strong variability of Z-sources, extracting power-color and total RMS variability features is challenging without proper segmentation of stationary intervals and a careful choice of integration time that ensures the selected interval satisfies the required SNR threshold. Thus, we began considering the Atoll sources only in our dataset, which are relatively less variable during the outburst phase.

For Atoll sources in the HID plot, during an outburst, the source covers a 'C'-shaped track. Different regions of this track correspond to different spectral states. Sources are observed in the upper and lower banana regions when they are in High/Soft State. Similarly, the source is seen in the Island and Extreme Island regions of the plot when it is in the Low/Hard State. Thus, the Atoll sources are further classified into four classes based on spectral states, which are

1. Upper Banana
2. Lower Banana
3. Island
4. Extreme Island

- *It is important to note that the nomenclature used for labeling corresponds to the regions in the HID/CCD projections, which are* ***uniquely*** *associated with the different spectral states that an Atoll source moves through during an outburst.*

Additionally, to obtain the spectral state labels, we referred to previous works that performed detailed spectral analysis and reported information related to the spectral state for each observation. Table 3.4 lists the observations, inferred states, and corresponding references. Observations

with available labels from studies based on AstroSat data were included in our dataset labeling.

The labeling was carried out based on the following conditions:

1. If a paper explicitly reported the spectral state for a given observation, we directly adopted that label.

2. In cases where different studies reported conflicting states for the same observation, we used our HID, CCD, and PCD projections of sample space to assign the label. In our analysis, we include large sample of data which corresponds to have full evolution tracks in the HID, CCD and PCD projections. As a result, for conflicting reports related to same observation, the position of a given observation can be checked, allowing a consistent label to be inferred, in agreement with at least one reported classification.

3. Since many studies infer states relative to other observations of the same source (e.g., one observation being softer or harder than another), relative classification can sometimes be ambiguous. In such cases, we adjusted the labels based on the overall track structure seen in the HID, CCD, and PCD projections in our dataset.

4. For some black-and-white cases, where the state is very clearly defined, the classification was straightforward. These cases lie in extreme regions of the HID, CCD, and PCD projections in our dataset and can be clearly assigned to a particular state. Such cases were also cross-checked with the full track evolution to ensure consistency.

Table 3.4: List of observations along with inferred spectral states and corresponding references.

Sr. No.	Source Name	Obs. Date	Inferred State	Reference
1	4U 1608−52	28-08-2016	Upper Banana	(1)
		14-07-2020	Upper Banana	
2	4U 1636−536	15-02-2016	Island	(2), (3), (4),
		02-07-2016	Island	(5)
		13-08-2016	Lower Banana	
		09-05-2018	Island	
		19-08-2018	Island	

Continued on next page

Table 3.4 continued

Sr. No.	Source Name	Obs. Date	Inferred State	Reference
		18-08-2018	Island	
		28-02-2017	Island	
3	4U 1702−429	27-04-2018	*Island*†	(6)
		08-08-2019	Extreme *Island*†	
4	4U 1728−34	07-03-2016	Upper Banana	(7), (8)
		19-02-2018	Upper Banana	
		17-07-2018	Upper Banana	
5	4U 1735−44	13-08-2017	Lower Banana	(9)
		31-07-2018	Lower Banana	
		24-08-2021	Lower Banana	
6	GS 1826−238	09-03-2016	Upper Banana	(10)
		17-08-2016	Lower Banana	
7	4U 1820−30	15-04-2022	Upper Banana	(11)
		10-03-2017	Upper Banana*	
		15-04-2022	Upper Banana*	
		26-05-2017	Upper Banana*	
8	XTE J1701−407	19-08-2017	Extreme Island*	
9	4U 1812−12	26-07-2018	Extreme Island*	
10	GX 3+1	29-04-2018	Upper Banana	(12)

† - Above condition 3 is applied for the labeling.

* - Above condition 4 is applied for the labeling.

References:

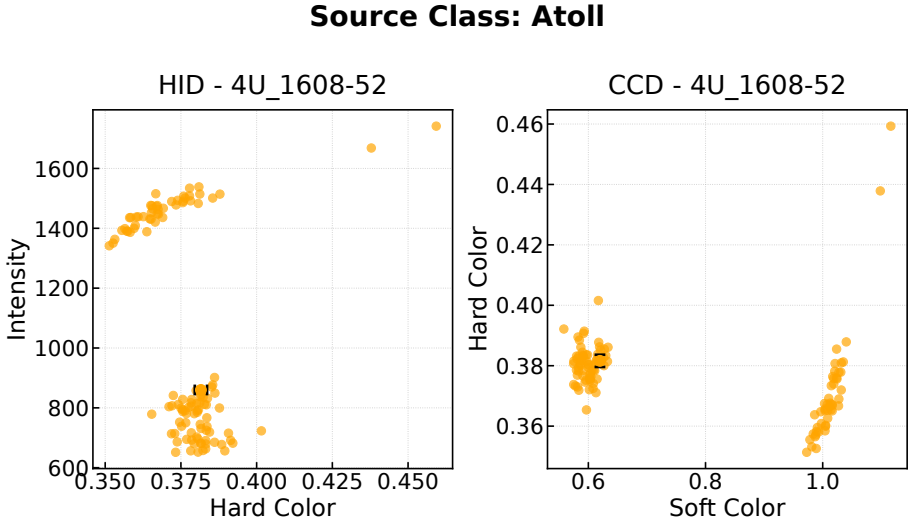
(1) Bhattacherjee et al. (2024b); (2) Beri et al. (2018); (3) Chattopadhyay et al. (2025); (4) Roy et al. (2022); (5) Kashyap et al. (2021); (6) Varun et al. (2024); (7) Bhattacharyya et al. (2018); (8) Salgundi et al. (2024); (9) Lavanya et al. (2024b); (10) Agrawal et al. (2022); (11) Pradhan et al. (2023); (12) Thomas et al. (2023b)

3.5 Feature Visualization

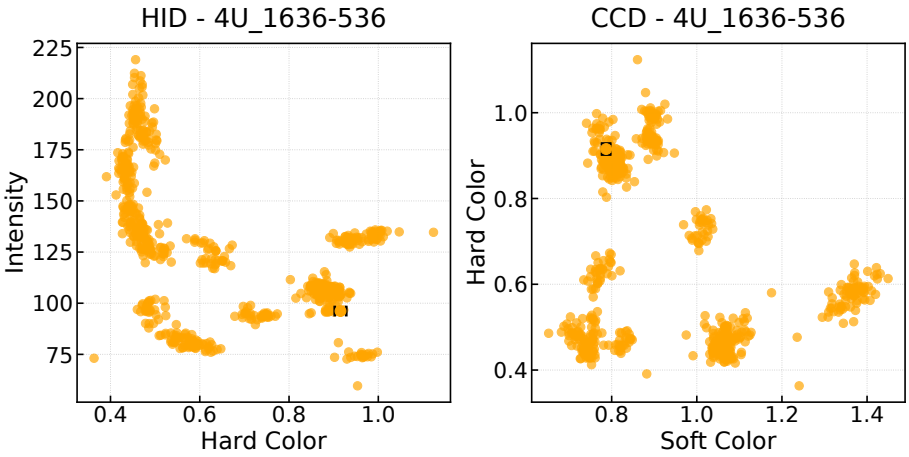
After the feature sample space is populated, we present different projections of the dataset to understand how well the data separates by source class and corresponding spectral states (for Atoll sources).

3.5.1 Hardness-Intensity Diagram and Color-Color Diagram

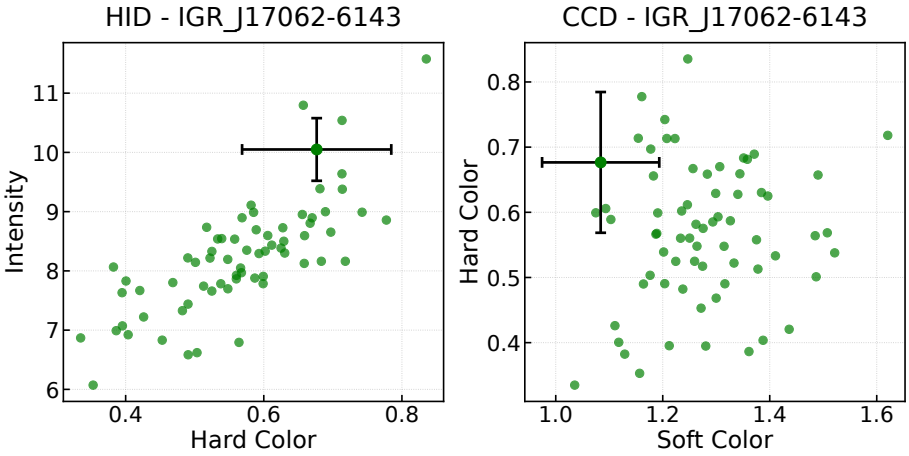
The Hardness-Intensity diagram (HID) is one of the most widely used tools for studying the spectral behavior of X-ray binaries. It provides a variation of total intensity with the hard color, indicating the source's state evolution. The HIDs were first generated for individual sources and then combined and color-coded into a single plot to provide a collective view of how different classes occupy distinct regions of feature space. Similar to HID, color-color diagrams provide an alternative projection of the feature sample space by comparing hard and soft color values. Fig. 3.3 and fig. 3.4 show the HID and CCD plots for selected sources and the highlighted source classes for the complete dataset.



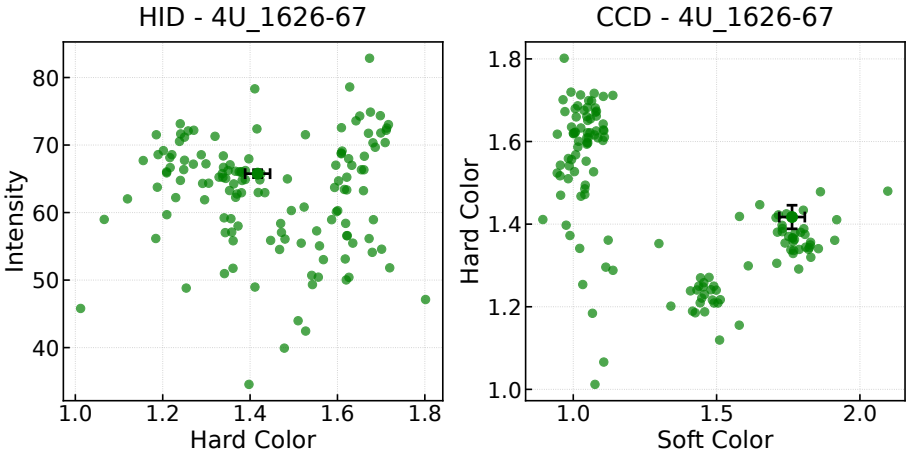
Source Class: Atoll



Source Class: Pulsar



Source Class: Pulsar



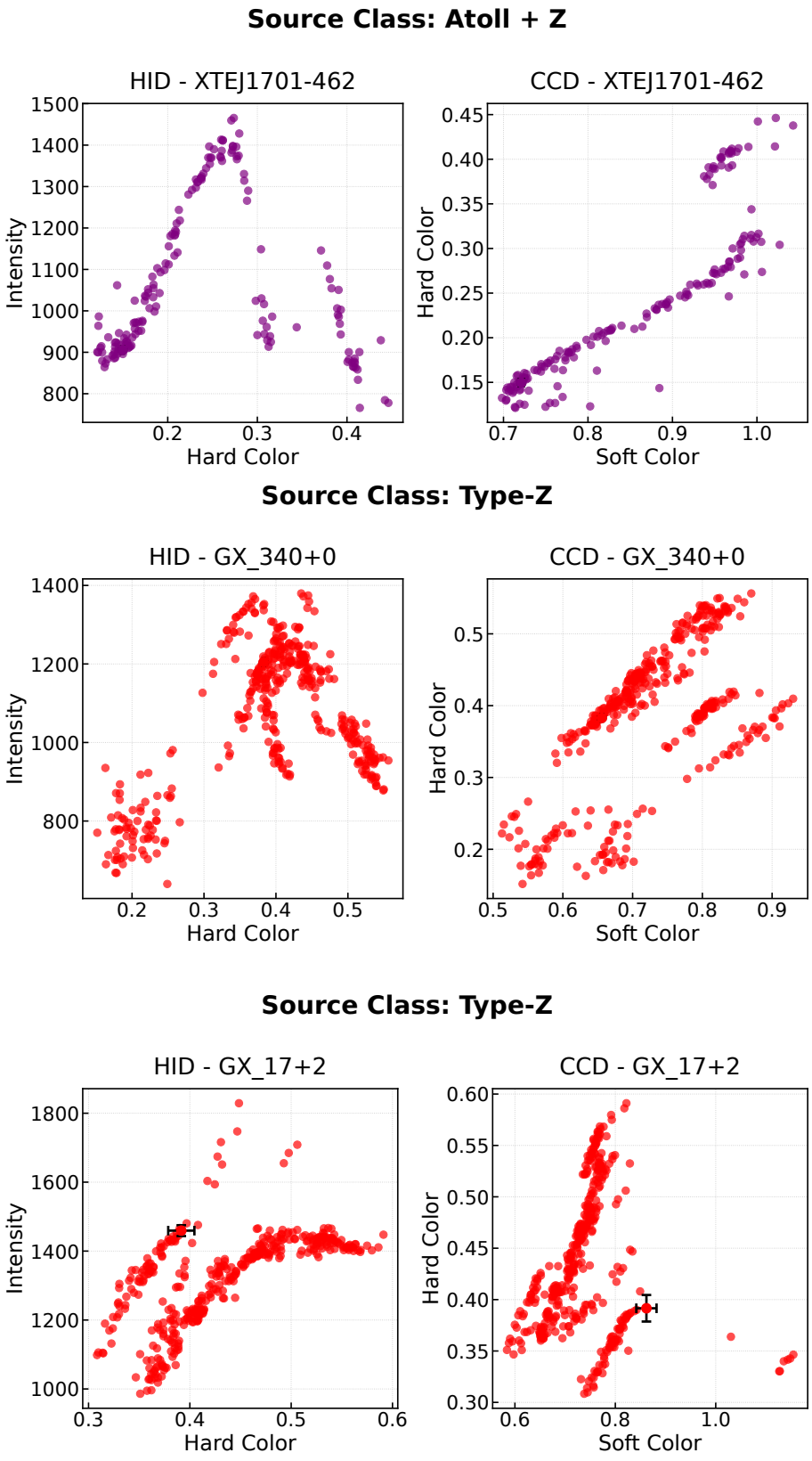
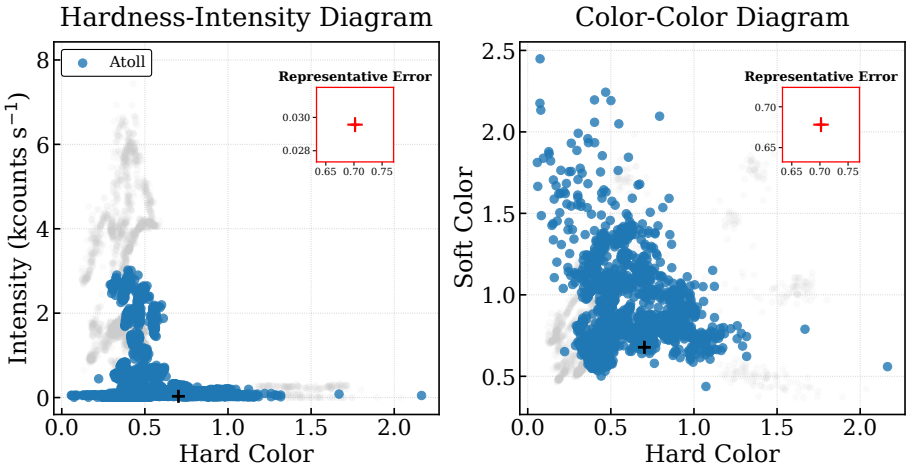
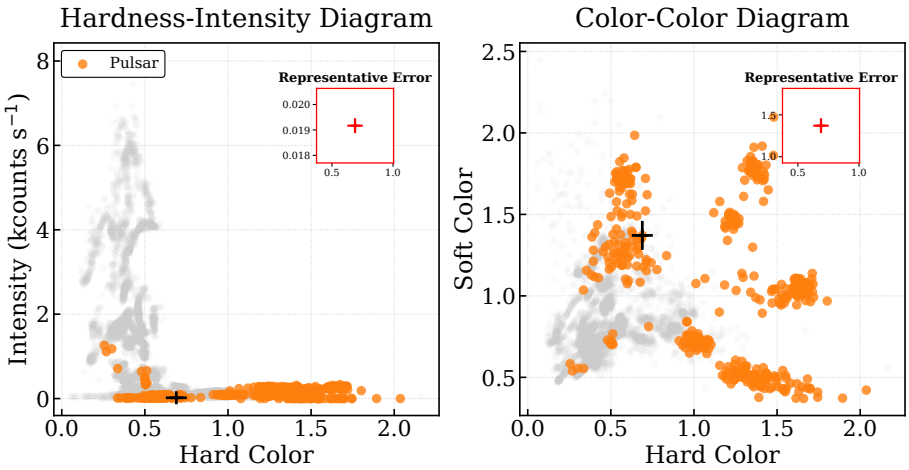


Figure 3.3: Hardness–Intensity Diagram (HID; left) and Colour–Colour Diagram (CCD; right) for selected sources. A representative error bar (black) is shown for a single data point for clarity; error bars for all points are omitted to avoid visual clutter.

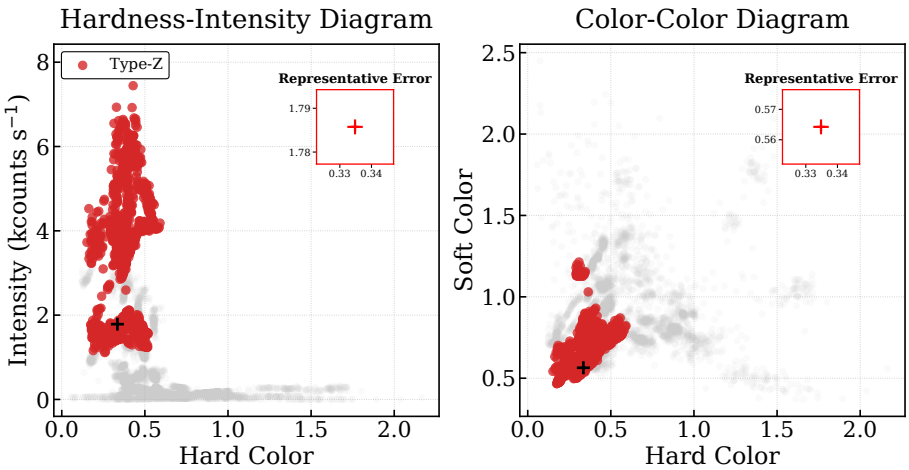
Source Class: Atoll



Source Class: Pulsar



Source Class: Type-Z



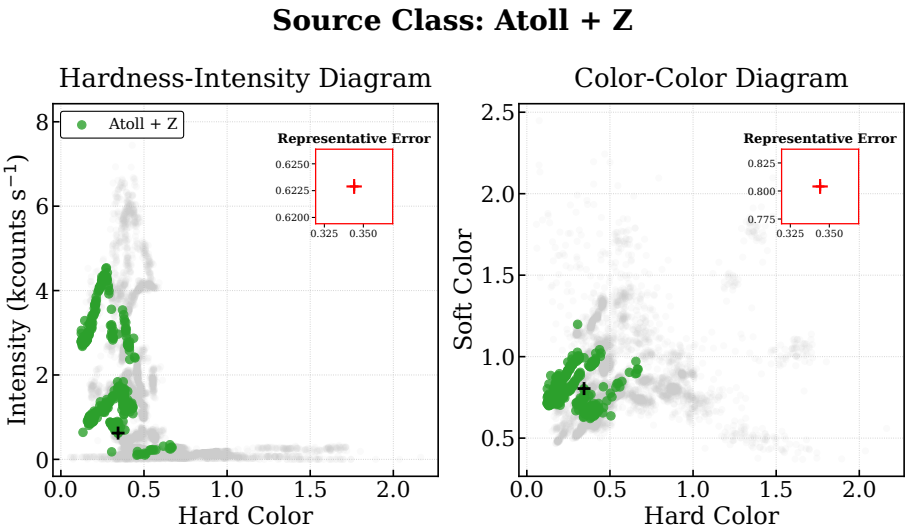


Figure 3.4: Combined HID and CCD showing all sources in gray. The highlighted source class is marked to indicate its position within the overall feature dataset.

3.5.2 Color-Color-Intensity Diagram

Vrtilek & Boroson (2013) introduced Color-Color-Intensity (CCI) plots to visualize spectral and intensity variations in X-ray sources. The CCI diagram is a 3D plot with hardness ratios along two axes and relative intensity as the third dimension.

In Fig. 3.5, the CCI plot is presented for all observations from three different viewing angles, with the points color-coded according to their respective source classes.

In Fig. 3.6, each source class is highlighted on the 3D CCI plot to track its evolution in the entire feature sample space.

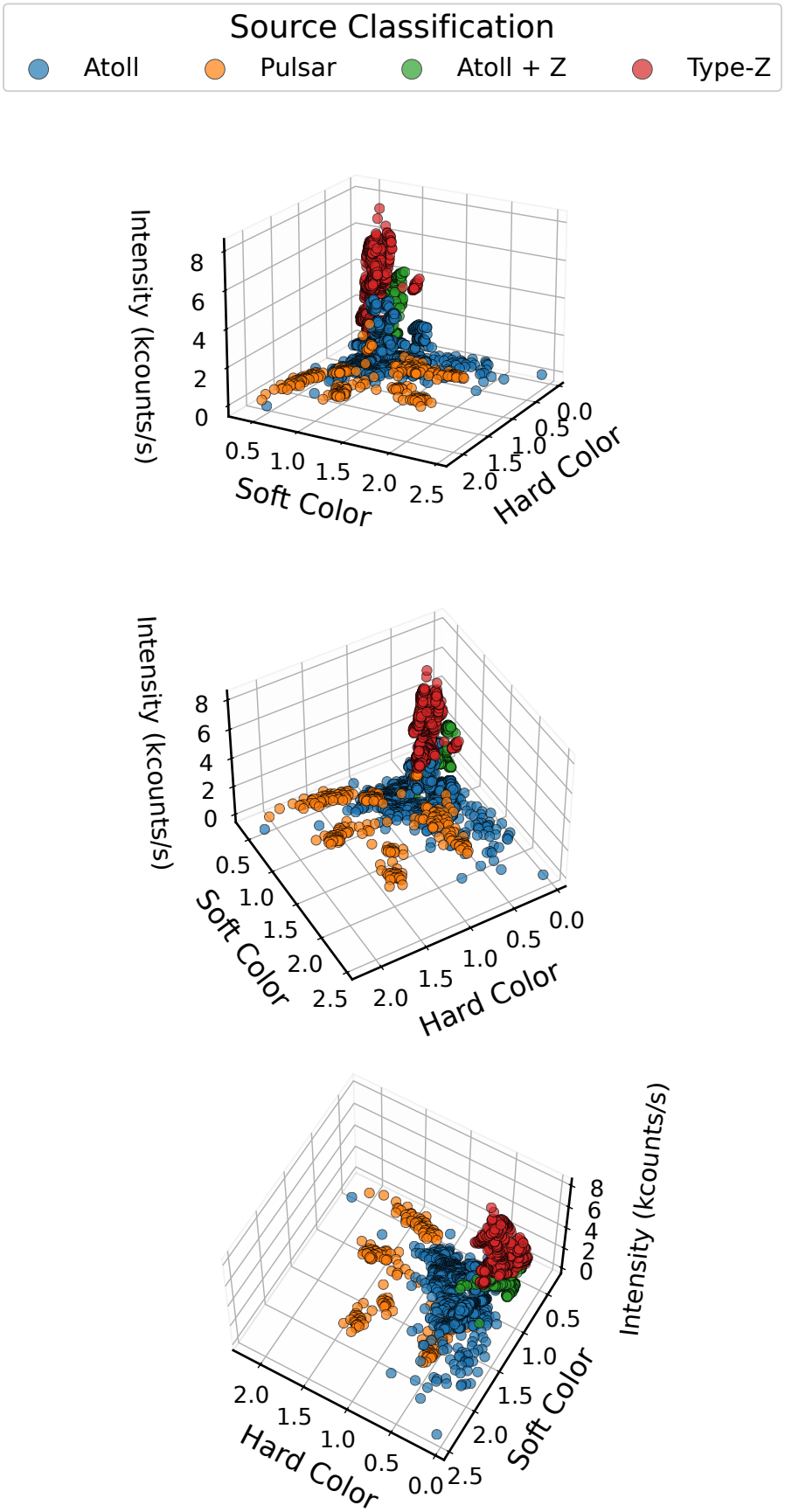


Figure 3.5: Three-dimensional Color–Color–Intensity (CCI) diagram visualized from different viewing orientations. The CCI is defined in terms of hard color, soft color, and normalized intensity along three axes.

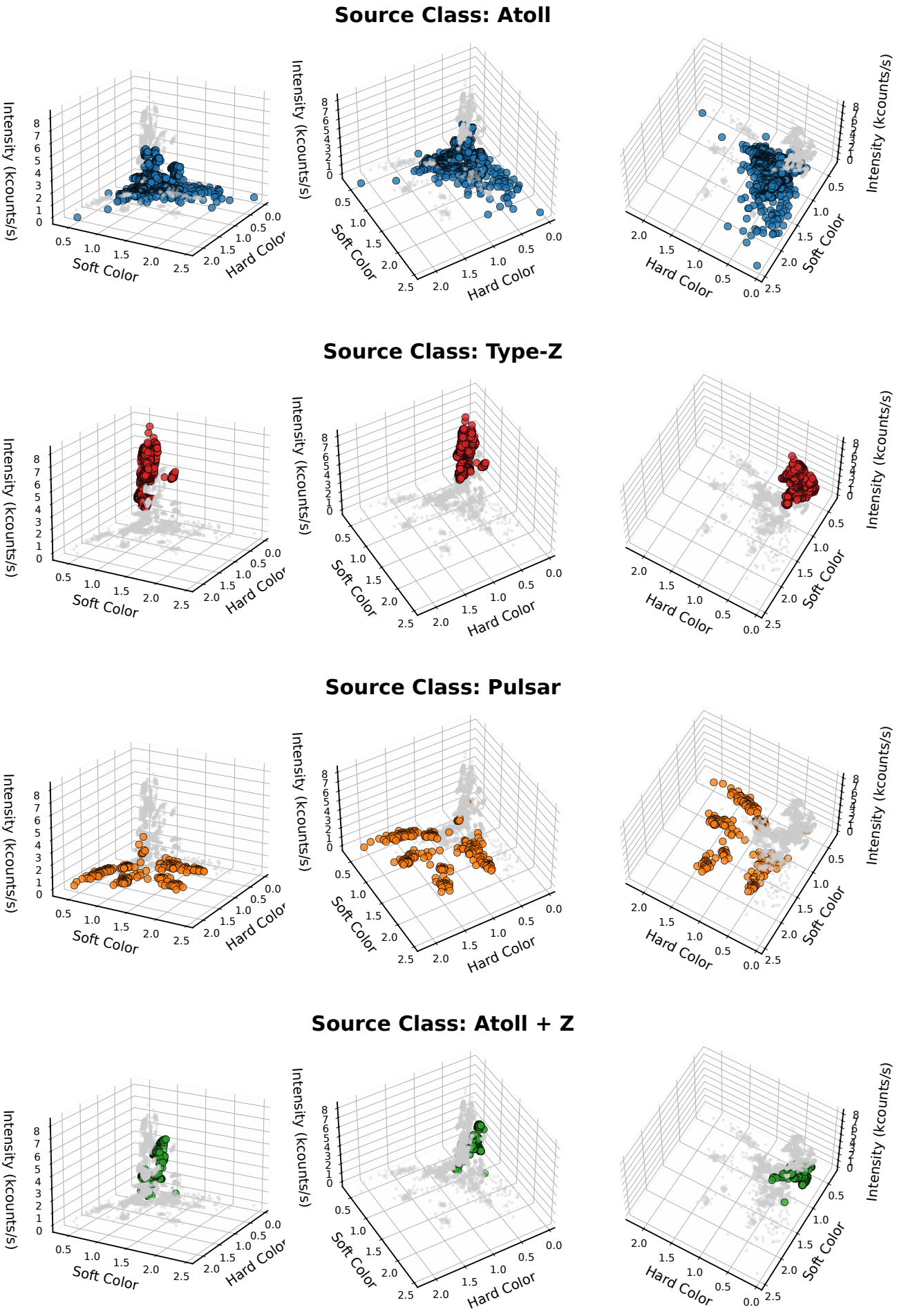


Figure 3.6: Color–Color–Intensity (CCI) diagrams for different source classes (Atoll, Z, Pulsar, and Atoll+Z). The highlighted points correspond to the selected class, while the grey points represent all other data.

3.5.3 Power Color Diagram (PCD)

The Power Color Diagram (PCD) provides a projection of power-color values onto the feature space. Fig 3.7 shows the PCD plots for the Atoll sources, with the labeled states color-coded.

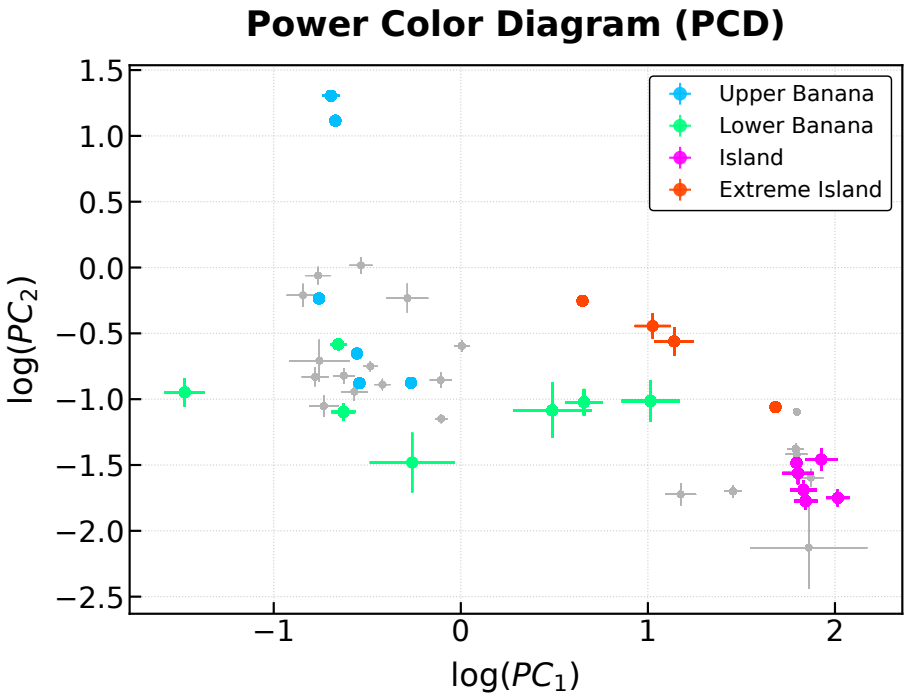


Figure 3.7: Power colour diagram plotted in $\log_{10}$ scale for each power color value. Data points are color-coded by spectral state, while grey points represent unlabeled data.

So far in this chapter, we have discussed source selection, the feature extraction process, data set labeling, and visualized the feature sample space for the source classes. In the upcoming chapter, we first discuss the basic theory underlying different models employed for the classification tasks, then present the methodology adopted for classification and model selection.

Chapter 4

Methodology

In this chapter, we present the machine learning techniques used for source and state classification, describe the methodology for preparing the training dataset, and detail the hyperparameter tuning performed to optimize model performance.

4.1 Classification Algorithm

Based on the analysis of previous works (see Sec. 1.5.1) and the nature of our labeled dataset, we adopt four machine learning algorithms for the classification tasks:

1. K-Nearest Neighbors (KNN)
2. Support Vector Machines (SVM)
3. Random Forest (RF)
4. XGBoost (eXtreme Gradient Boosting)

In this section, we outline the basic framework of these algorithms.

4.1.1 K-Nearest Neighbors

KNN method (Cover & Hart, 1967) is widely used, owing to its effectiveness and proven performance in similar classification tasks. For a test data point, this method assigns a label by calculating the K nearest neighbors, then assigning the label of the majority class among them. The value of K is user-specific and depends on the dataset's requirements. Given below is a detailed flowchart and logic of how classification is carried out using the KNN method.

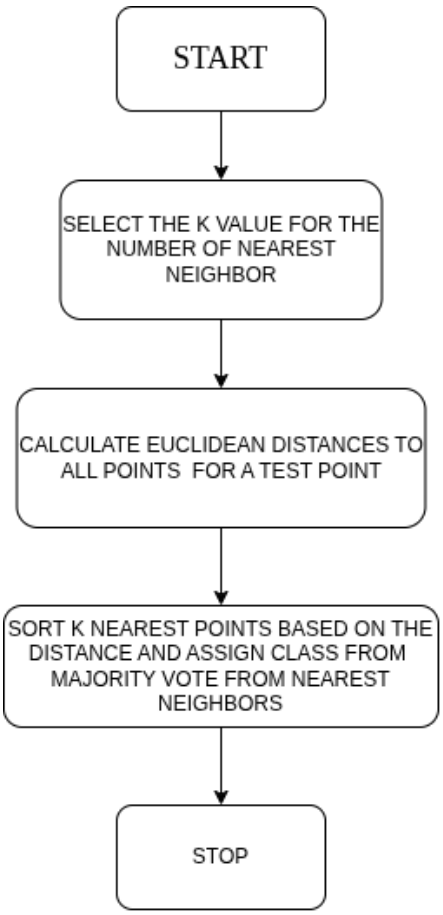


Figure 4.1: A Flowchart of the K-nearest neighbor algorithm is shown. Once the K value is chosen, the algorithm calculates distances from a single test point to all points in the training dataset, then selects the class with the majority count among the nearest neighbors. The process is repeated for multiple test data points.

4.1.2 Support Vector Machine

Support Vector Machine (Vapnik, 1995) is one of the most popular supervised learning methods. The effectiveness arises from its ability to classify complex data (linear and nonlinearly separable) efficiently and easily. It is reported that SVM performs well even in higher-dimensional spaces.

Given a feature space, Support Vector Machines (SVMs) aim to identify the optimal hyperplane (decision boundary) that maximizes the separation between data points of different classes. This is achieved by maximizing the margin, defined as the distance between the hyperplane and the nearest data points from each class (support vectors), while ensuring correct classification of the training data.

Fig 4.2 gives a schematic of hyperplane margin optimization.

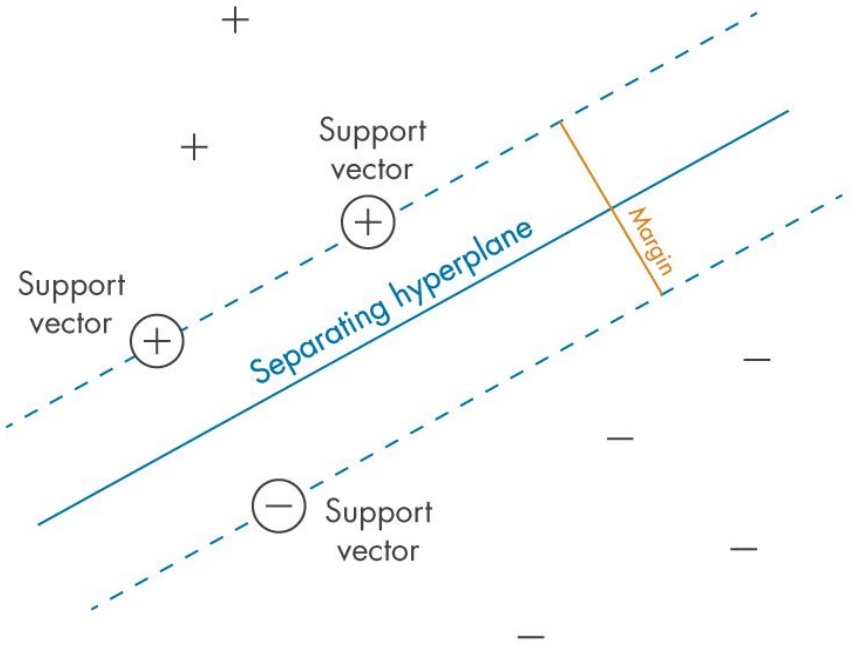


Figure 4.2: The figure provides a schematic for the SVM algorithm for the margin width maximization. Here, the width is maximized so that data points from two classes (represented by plus and minus signs in the figure) are well separated. [Image reference: https://uk.mathworks.com/discovery/support-vector-machine.html]

For a test data point, a vector is drawn considering a reference point as the origin; this vector is known as the Support Vector. In SVM, the Support Vector is projected onto the vector joining the origin to the hyperplane, which is also perpendicular to the hyperplane. The class of a data point is determined by the sign of its projection onto the normal vector of the hyperplane. In other words, the decision function determines which side of the hyperplane the data point lies on, thereby assigning it a class label.

For linearly separable data, identifying the hyperplane is straightforward. But for a nonlinear, complex dataset, identifying a hyperplane is difficult. To address this difficulty, SVM uses the 'Kernel Trick'. The basic idea behind this method is to project the data to higher dimensions where the separation can be easily achieved. Kernel functions compute the projection of a data point into a higher-dimensional space, providing a classification indication for that point.

4.1.3 Random Forest (RF)

Random Forest (Breiman, 2001) is a supervised machine learning algorithm that combines predictions from multiple decision trees and determines the final output by majority voting.

Decision Trees are supervised learning models that perform classification by repeatedly partitioning the feature space into distinct regions. At each node, the data are split based on a feature and a threshold chosen to maximize the separation between classes. This process continues until a stopping condition is met, resulting in a tree-like structure where each leaf node represents a class label. Due to their hierarchical structure, decision trees are simple to interpret and can capture non-linear relationships in the data.

The RF algorithm begins by creating multiple samples from the original feature dataset. In this sampling method, the data are selected with replacement, meaning the same data points can appear multiple times in a sample. This process is known as 'Bootstrap' or 'Bagging'. The decision trees are trained on each sample, and the majority vote determines the classification of data points. Fig 4.3 provides the schematic flowchart of RF working.

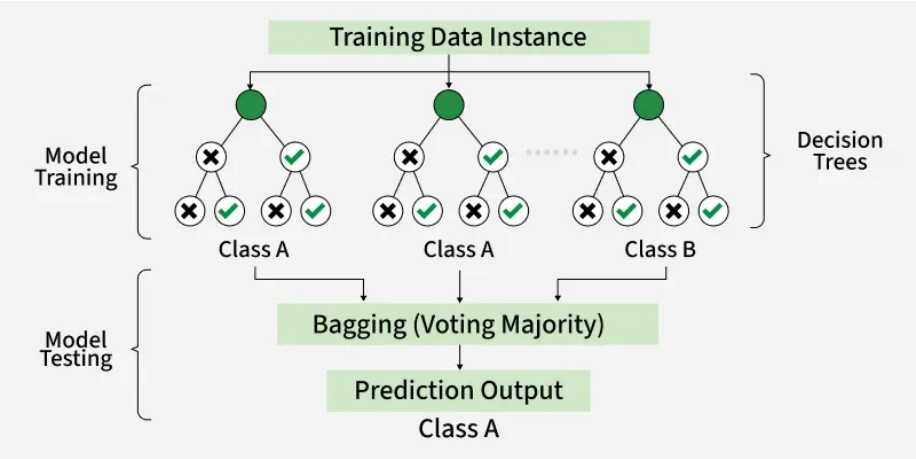


Figure 4.3: Schematic representation of the Random Forest (RF) algorithm, illustrating the bagging approach used to train multiple decision trees. The final prediction is obtained through majority voting among the individual tree outputs. [Image reference: https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/]

4.1.4 XGBoost (eXtreme Gradient Boosting)

XGBoost is an ensemble method (using multiple weak models to obtain strong results) that is highly effective and efficient for accurate classification of complex datasets.

Unlike Random Forest, where multiple decision trees were trained, and the majority vote determined the model prediction, XGBoost builds trees one after the other. Each new tree is trained to model (learn) the residuals from the previous tree, which boosts learning performance. The final result is a weighted sum of predictions from all the decision trees. It uses gradient descent to minimize the loss function and quickly learn from past mistakes. It is also reported that the model is efficient at classification even with missing data.

4.2 Training dataset preparation

The dataset comprises 6,018 data points from 135 observations across 35 sources for source-class classification, and 2,611 data points from 44 observations across 13 sources for state classification for atoll sources. For the state classification of Atolls, out of 2,611 data points, 1,477 are labeled.

The data points are not evenly distributed among all the sources. Thus, to ensure equal representation across all source classes, the training dataset should be carefully sampled. Bias can be introduced into the training dataset by adding more data points from a well-sampled source, and the algorithm will be trained more on those data points.

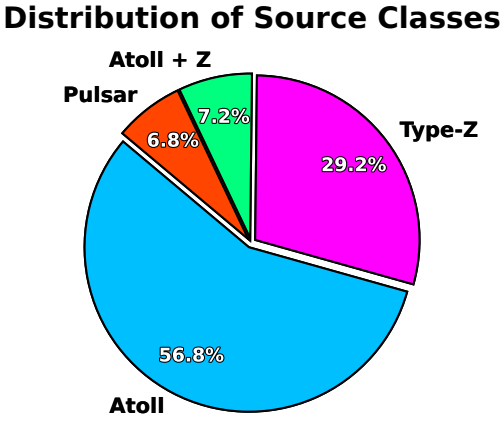


Fig. 4.4 illustrates the distribution of different classes in the feature space, clearly highlighting the presence of class imbalance. To address

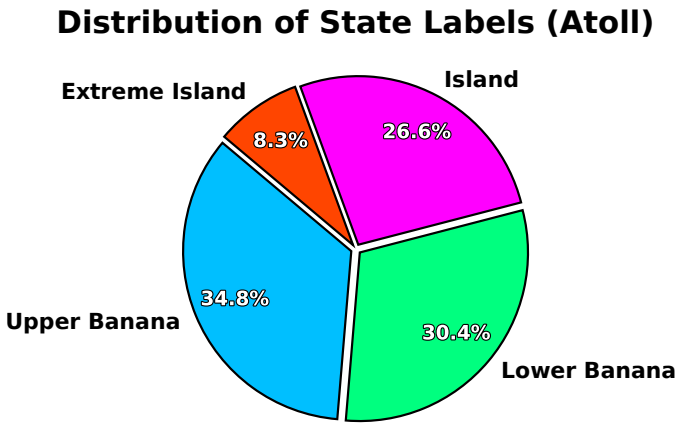


Figure 4.4: Distribution of dataset labels: The top panel (previous page) displays class labels, while this panel illustrates atoll state labels.

this imbalance in the dataset and avoid bias from multiple observations of the same source dominating the training, we adopt a stratified group-based split rather than manually assigning weights and constructing the training set. We used a Stratified Group K-fold splitting strategy, which takes care of two important things:

1. Stratified splitting ensures that the class proportions are maintained across all folds, which helps to make sure that all classes are represented in both the training and test samples.

2. Group splitting ensures that all observation information belonging to a class is entirely included within the same fold. This prevents any data leakage, which means avoiding training and testing on the same samples.

This ensures that the model is tested on unseen data and no data leakage or bias is introduced, leading to better, more reliable, and more robust classification.

We employ a 5-fold splitting scheme, in which the dataset is partitioned into 5 subsets: 4 for training and 1 for validation. This corresponds to an approximate 80–20 train-test split. We found that the splitting scheme is sensitive to the random seed, as the folds are generated by randomly selecting data points and then optimizing for class imbalance. Thus, to obtain efficient and robust classification, the model is evaluated across multiple training iterations, so that each data point is tested at least once.

4.3 Cross Validation and Model Hyperparameter Tuning

Having introduced the machine learning algorithms, we now focus on selecting the best values for hyperparameters (model-specific variables that are defined initially before the training process and are used to tune the performance of the model) for each model. Hyperparameter tuning is done using Cross-Validation across multiple combinations of variables, then selecting the optimal values that yield the best model performance.

The selection of hyperparameter tuning depends on multiple factors (performance scores), which are explained as:

- **Accuracy**: Accuracy is calculated as the ratio of the correct predictions to the total predictions.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{4.1}$$

  where TP, TN, FP, and FN refer to True Positive, True Negative, False Positive, and False Negative, respectively. Here, a true positive means that the sample actually belongs to that class and is correctly predicted. Whereas false positives and false negatives denote the incorrectly predicted or missed samples for that class.

- **F1 score**: F1 score provides a balanced measure of correctness, combining precision and recall. For a given class, precision and recall are given by eq 4.2

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \end{aligned} \tag{4.2}$$

  The corresponding F1 score is given by the harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.3}$$

  For imbalanced datasets, the F1-score is preferred over accuracy, as accuracy may be dominated by the majority class. The F1-score incorporates both precision and recall, providing a more balanced evaluation of model performance across all classes.

- **Cohen's Kappa**: This metric evaluates model performance by measuring the agreement between predicted and true labels. It compares the observed accuracy with the expected accuracy, where the expected value is calculated using the class proportions of the true and predicted labels, assuming random predictions. This helps account for agreement occurring by chance and provides a more reliable metric, especially for imbalanced datasets.

  For a value $\kappa > 0.6$, it is supposed to be a reliable model, and lower values suggest that the correctness in the classification is obtained by luck or random guessing.

We perform a detailed performance analysis using cross-validation with an observation-based 5-fold split. Additionally, we employ a leave-one-source-out cross-validation strategy to further evaluate model robustness.

We find from the comparison analysis among multiple algorithms (refer to section 5.3) that, among the four algorithms discussed in 4.1

1. For **source-class classification**, the **K-Nearest Neighbors** algorithm performs the best.

2. For **state classification** in Atoll sources, the **Support Vector Machine** performs the best.

Below, we present the results from cross-validation used to tune the hyperparameters for the best-performing models.

4.3.1 Source-class Classification: K-Nearest Neighbors

We performed hyperparameter tuning to determine the optimal value of K, representing the number of nearest neighbors, using cross-validation. This is carried out using the `StratifiedGroupKFold` split function from the `sklearn` library, which ensures that class proportions are preserved across folds while keeping observations from the same source within a single fold. Given the data imbalance, we employed the `UnderSampler` library from the `imblearn` module to ensure that, within a single split, data points from all classes are equally represented across the folds. Undersampling is a common sampling technique employed to ensure that minority classes in the feature dataset are well represented.

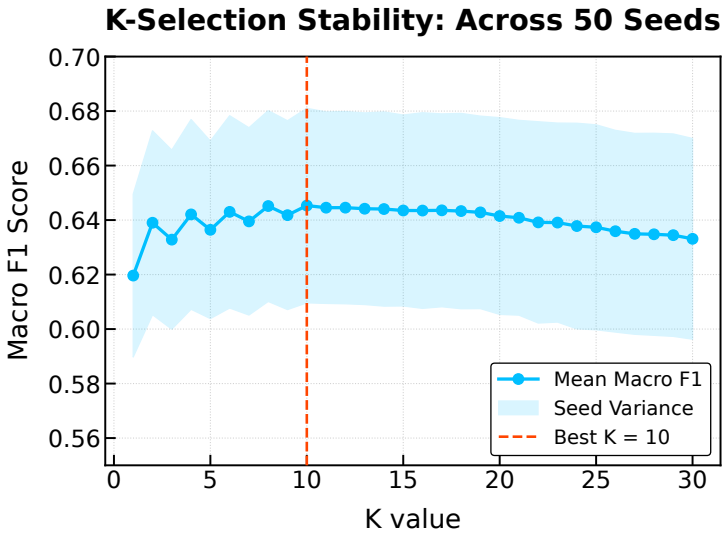


Figure 4.5: The plot of K values vs the F1 score is given, where the scores are an aggregate of random 50 seed values. For K = 10, the model performs better compared to other values of K.

We performed a stability analysis of the undersampling logic, as it is highly sensitive to the random seed. Thus, we found that for 50 seeds, results are stable and robust, with high model performance. For hyperparameter tuning, performance is evaluated across different values of K, and the value yielding the best average cross-validation score is selected as the optimal choice. Fig 4.5 refers to the plot where, for different K values, the model is evaluated, and results are shown as average performance F1 score for 50 such realizations.

Thus, for K = 10, the performance metrics attain their maximum values, and so it is chosen as the optimal number of neighbors. In addition to the number of neighbors, the choice of weighing scheme and distance metric is also an important hyperparameter in the KNN method. Since this supervised method depends on the distribution of data points in the sample space, the Distance metric is usually chosen `'Euclidean'`, and the Weightage parameter is based on the `'Distance'`. This means that greater weight will be assigned to points closest to the test point, even within the selected K neighbors used for majority voting. For our analysis, we proceed with the same choice of schemes.

Leave-one-out cross-validation was also used to determine the best K value. In this approach, the model is trained on all sources except the one it is tested on. This process is repeated for all sources to ensure that each has served as the testing dataset. In the upcoming chapter, we discuss

in detail the accuracy scores obtained for each source when left out of the training data during testing. Such a validation method helps evaluate how well the model performs when presented with a completely unseen source for prediction.

4.3.2 State Classification for Atolls: Support Vector Machine

Similar to the KNN hyperparameter tuning approach, for state classification, due to data imbalance, `Undersampler` is used. For SVM, the hyperparameters are discussed as follows:

- C (Regularization parameter): This parameter controls the tradeoff between maximizing the width of the hyperplane and minimizing the training errors.
- Kernel: This parameter states the function that can be used to project the data points from a complex sample space to higher dimensions, where easy classification between classes can be obtained.
- Gamma (Kernel Coefficient): Balances between overfitting and underfitting the decision boundaries.

Usually, for tuning purposes, the C value ranges from 0.1 to 100, whereas the γ value ranges from 0.001 to 1. The variation in the model's performance with C is observed to follow a logarithmic scale. For kernel selection, we tried multiple functions (Radial Basis Function, Linear, Nonlinear, and polynomial), and Radial Basis Function produced the best results. We used the `GridSearchCV` module from `sklearn` for hyperparameter tuning, searching over C values in [0.1, 1, 10, 100] and γ values in [0.001, 0.01, 0.1, 1]. Table 4.1 provides the results from this grid search method, where for every combination of the above-mentioned parameters, the model's performance is evaluated.

Thus, for C = 100 and γ = 0.01, the model's performance metrics are the best. From the table, it can be observed that as C increases from 10 to 100, the metric scores improve, suggesting that the model penalizes misclassifications more for higher C values and thus fits the training data more closely. Also, for γ, a very small value may under-fit the dataset, and a very high value can sometimes over-fit the data. Thus, we finalized the model hyperparameters as C = 100 and γ = 0.01 for the SVM with the RBF kernel.

Table 4.1: Grid search results for SVM hyperparameters for rbf kernel.

C	γ	F1-score
0.1	0.001	0.4460
0.1	0.01	0.7799
0.1	0.1	0.8036
0.1	1	0.5808
1	0.001	0.7689
1	0.01	0.8147
1	0.1	0.8710
1	1	0.5220
10	0.001	0.8143
10	0.01	0.8835
10	0.1	0.8781
10	1	0.5213
100	0.001	0.8733
100	0.01	**0.9055**
100	0.1	0.8781
100	1	0.5213

So far, we have discussed how the feature sample space is constructed and how hyperparameter tuning is performed for different supervised algorithms in both source-class and state-classification settings. We now wanted to examine how an **unsupervised** algorithm operates on these unlabeled datasets. We especially wanted to investigate whether an unsupervised algorithm could reproduce state classification results consistent with those obtained from supervised methods.

This analysis provides insight into the statistical distribution of the features. It also helps us probe whether the features themselves are sufficient to naturally separate the classes into distinct clusters, or if any additional clustering is needed. These additional clusters may represent unrecognized states or transitional regions between them, as the source evolution is a continuous process. In the following section, we present the hyperparameter tuning carried out for an unsupervised algorithm.

4.4 Unsupervised Machine Learning Method

The primary motivation for this clustering task is to capture the complete evolution of a source during its outburst, including the intervals in which it transitions from one X-ray state to another. We are further motivated by the fact that, in the feature sample space, we have an entire evolutionary track of atoll sources when combined across multiple dimensions. As

the tracks are continuous, we expect that the feature not only contributes to classifying the states, but also the intermediate transitional regions between them.

Such transitional regions are very important to study, as they help to probe into sudden changes in the source's physical properties, such as mass accretion rate, disk-corona geometry, and jet activity.

4.4.1 Clustering Algorithm

To identify a robust and correct clustering, a soft probability-based cluster assignment is a suitable starting point for this analysis, compared to a hard assignment. The benefit of using a soft probability-based assignment is that it indicates the likelihood of each data point belonging to each cluster. Owing to this reason, we began our clustering process using the Gaussian Mixture Model (GMM).

The Gaussian Mixture Model assumes that a complex dataset can be approximated as a mixture of multiple Gaussian distributions, with each distribution representing a distinct cluster in the data. Each Gaussian component is characterized by its mean, covariance matrix, and the weight assigned to that component. The weight determines the proportion of data represented by that component.

In this case, directly maximizing the log-likelihood of a multiple-Gaussian (GMM) becomes difficult due to a mathematical challenge. If we already know which points belong to which Gaussian component, the component's parameters (mean, covariance, and weight) can be easily calculated. Conversely, if we know the component's parameters, assigning each point to its most likely Gaussian would be easy. However, both these quantities are unknown simultaneously.

This interdependence makes log-likelihood estimation mathematically difficult, so the Expectation-Maximization algorithm (Dempster et al., 2018) is used to optimize the component parameters.

The algorithm flow is as follows:

1. Initialization: The process starts by random initializations for means, covariances, and weights for K Gaussian components. Since clustering is sensitive to initialization, multiple random initializations (controlled by the hyperparameter n_init) are used, and the best result is selected.

2. Expectation: Based on the current parameter values, the algorithm computes the probability that each data point belongs to each Gaussian component. These probabilities depend on how close the point is to the component's mean, the component's spread (covariance), and the component's weight in the mixture.

3. Maximization: The model updates the mean, covariance, and weights of the components based on the probabilities obtained in the above step. The updated parameters are chosen to increase the likelihood of the observed data under the mixture model.

4. Convergence: The Expectation-Maximization step is repeated iteratively. The algorithm stops when there is no significant improvement in the results from further updates.

The GMM model uses the following hyperparameters:

- `n_init`- Controls the number of initializations.
- `n_components`- Specifies the number of Gaussian components (clusters) to fit the data.

4.4.2 Hyperparameter Tuning

Since the Gaussian Mixture Model (GMM) relies on random initializations, its solution can vary with the seed value. To reduce this variability, clustering performance was evaluated across 50 seeds, and the mean and standard deviation of the results were used for comparison.

We explored models with 1 to 40 components. Very small values of $n_components$ may lead to underfitting, where the model is unable to capture the data's structure and cannot explain the labeled subset. Also, a very high value may lead to overfitting. Therefore, an intermediate value must be chosen. We explore the use of the Bayesian Information Criterion (BIC) for hyperparameter tuning. However, for a range of hyperparameters, it showcased a linear behavior, and no elbow point can be found. This makes BIC unsuitable for hyperparameter selection for the current study.

The ARI is a metric score that measures how well the predicted clusters reproduce the true structure: a value of 0 indicates that any agreement is purely by chance, whereas a value of 1 indicates perfect agreement between the predicted and true labels.

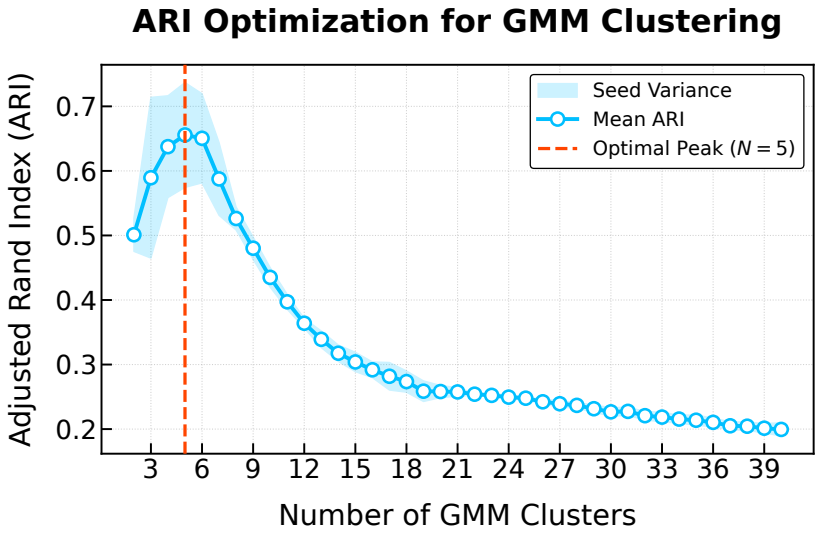


Figure 4.6: Adjusted Rand Index (ARI) as a function of the number of Gaussian components used in the GMM.

From Fig 4.6, it can be observed that the ARI results indicate that 5 components best explain the labeled data. We will explore the possible reasons behind this choice of hyperparameter value in the upcoming chapter.

Chapter 5

Results and Discussions

This chapter provides the detailed analyses of the results obtained using the machine learning methods described in Chapter 4.

5.1 Source-class Classification

We now present the results for the best-performing algorithms for source-class classification problems, along with the cross-validation results, in this section. The K-nearest Neighbor classifier performed best among the other adopted methods. Hence, using K=10, with `distance`-based weighting and the `Euclidean` distance metric (refer sec. 4.3.1), we performed an 80–20 train–test split. Hard color, soft color, and normalized intensity were used as features for classification. Fig 5.1 gives class-wise precision, recall, and F1-score.

Figure 5.2 provides the confusion matrix illustrating the classification performance across different source classes. The matrix is based on the mean of predictions over 50 random seeds. The performance scores obtained by the classifier are shown in Table 5.1.

Table 5.1: Overall classification performance metrics for the source-class classification using the KNN model. Errors represent the standard deviation (1σ) across all cross-validation folds.

Metric	Value
Overall Accuracy	$84.9 \pm 0.2\%$
F1-score	$64.5 \pm 0.5\%$
Cohen's Kappa (κ)	0.72 ± 0.03

Landis & Koch (1977) suggested that a Kappa value greater than 0.6 indicates substantial agreement, which means that the classification

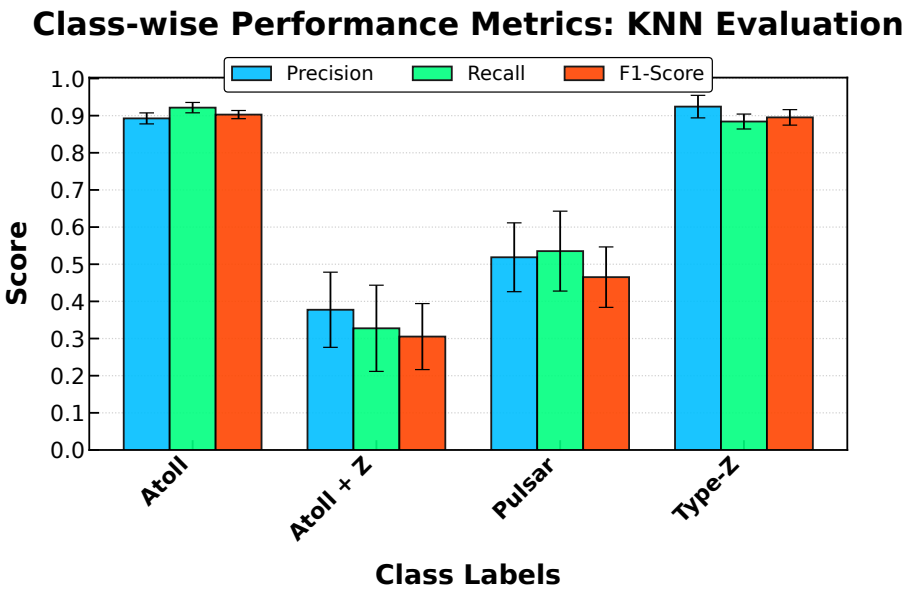


Figure 5.1: Figure shows class-wise classification performance scores (precision, recall, and F1-score, respectively) for source-class-based classification.

performance is significantly better than obtaining a good result just 'by chance'.

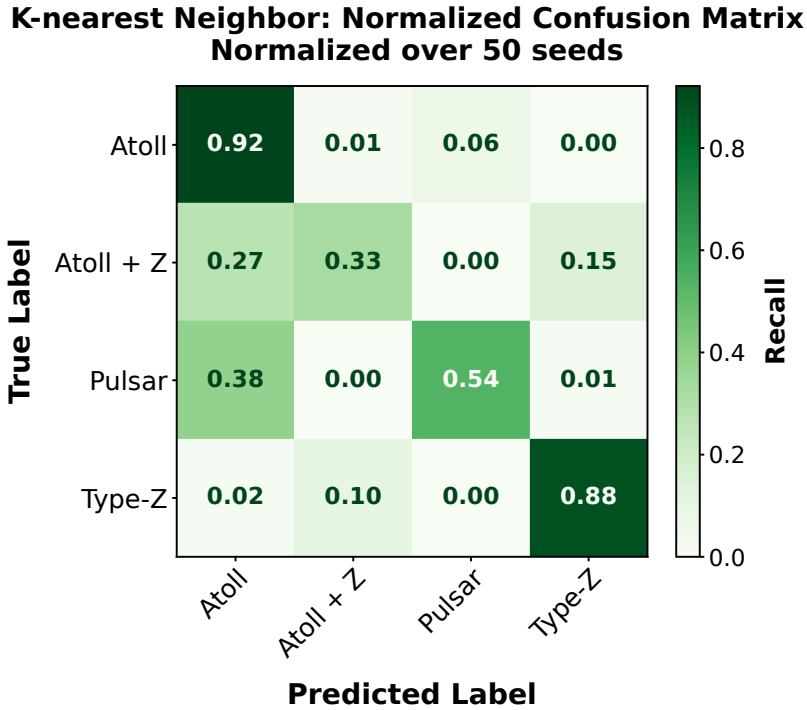


Figure 5.2: Confusion matrix for the KNN method used for source class classification.

From the class-wise metric scores and the confusion matrix, we can observe that the classifier struggles to achieve accurate classification for

the *Pulsar* and *Atoll + Type-Z* classes. For both classes, the F1-score is below the average (30% and 46.5%, respectively, for *Pulsar* and *Atoll + Type-Z*), and the confusion matrix shows the classifier is less confident in correctly classifying these classes, with confidence scores below 50%.

The comparatively poorer performance of the *Atoll + Type-Z* group is expected, as this class exhibits characteristics of both '*Atoll*' and '*Type-Z*' sources. Homan et al. (2010b) carried out a detailed study on one of the *Atoll + Type-Z* sources and showed that the transition occurs smoothly from *Type-Z* to *Atoll* behavior over the course of the outburst. This suggests an overlap in the two source classes. While some sources belonging to this class in our sample may exhibit steady evolution and non-transitional behavior across the *Atoll* and *Type-Z* classes during an observation. This leads the classifier to assign labels based on dominant behavior. Conversely, the presence of an intermediate transition region of *Atoll* and *Type-Z* behavior results in feature overlap. In both cases, the classifier may lead to misclassifications.

Straaten et al. (2005) find that the accreting millisecond X-ray pulsars showcase very similar timing properties to *Atoll* sources. Additionally, Gladstone et al. (2007b) conducted analyses of the spectral state evolution of *Atolls* and *Pulsars* (AMXPs) and concluded that both source classes broadly follow similar spectral properties as a function of mass accretion rate, with subtle differences. Owing to these similarities in the spectral and timing behavior of Atoll and Pulsar classes, the features can occupy overlapping regions of the feature space, leading to misclassification between the two classes.

Motivated by this intrinsic overlap and the need to assess how well the model generalizes across different sources, we adopt a Leave-One-Source-Out (LOSO) cross-validation approach. In each iteration, one source was completely excluded from the training set and then classified using the model trained on the remaining sources. This process provides the classification accuracy for the left-out source. The procedure was repeated for all sources using the analysis described above. Fig. 5.3 shows the classifier's performance for each source when treated as unseen data.

The results from Leave-One-Source-Out (LOSO) cross-validation are consistent with our above discussion of the *Atoll* and *Pulsar* misclassifications, as well as the *Atoll + Type-Z* confusion. Table 5.2 gives

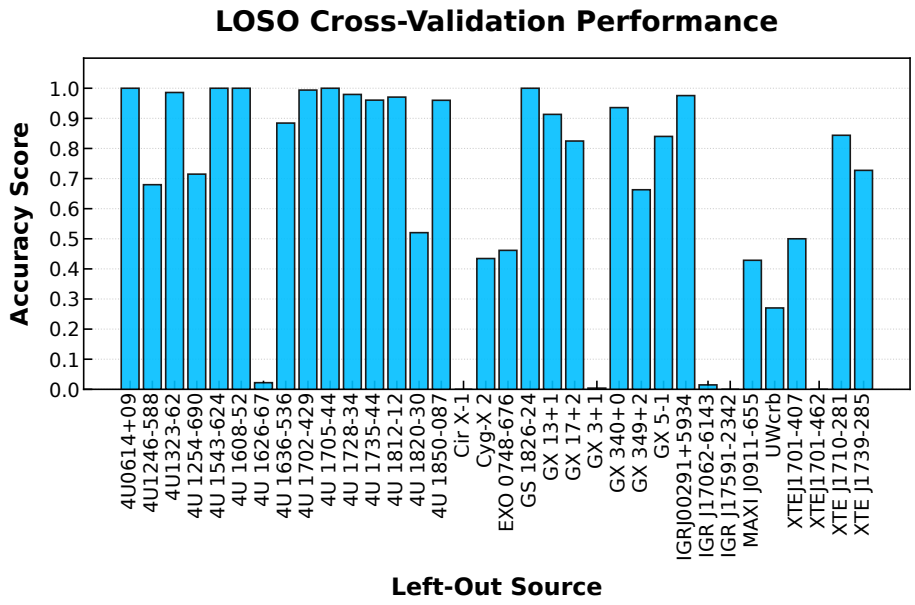


Figure 5.3: Accuracy of source classification under the Leave-One-Source-Out method.

Table 5.2: Summary of misclassified sources in source-class identification using Leave-One-Source-Out (LOSO) evaluation.

Source	True Class	Predicted Class	Classification Accuracy (%)
MAXI J0911−655	Pulsar	Atoll	43.9
Cyg X−2	Type-Z	Atoll + Z	42.6
UWcrb	Atoll	Pulsar	27.0
IGR J17062−6143	Pulsar	Atoll	29.4
4U 1626−67	Pulsar	Atoll	22.0
IGR J17591−2342	Pulsar	Atoll	0.0
GX 3+1	Atoll	Type-Z	0.0
Cir X−1	Atoll + Z	Type-Z	0.0
XTE J1701−462	Atoll + Z	Atoll	0.0

a summary of sources that are getting misclassified when we apply the KNN algorithm with the optimized hyperparameters to classify them.

It is interesting that for source GX 3+1, the classifier totally fails to predict the source class. **GX 3+1 is one of the brightest Atoll sources**. Along with sources such as GX 9+9, GX 9+1, and GX 13+1, it forms a class of bright Atoll sources (Hasinger & van der Klis, 1989). From the feature importance study, we found that **Intensity** is the **most significant feature** for source class-based classification, with an importance score of $\sim 0.35 \pm 0.004$, significantly higher than the color features (soft color: $\sim$ 0.20 ± 0.003, hard color: $\sim$ 0.18 ± 0.004). Because intensity is

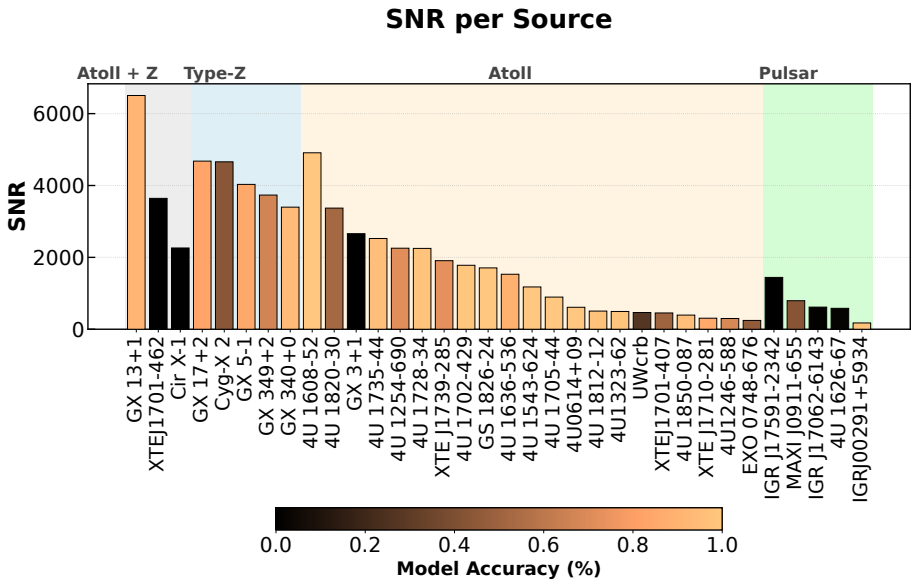


Figure 5.4: Average signal-to-noise ratio (SNR) distribution across different source classes, along with the corresponding classification accuracy. Pulsar sources exhibit comparatively lower SNR values, which correlate with reduced classification performance.

an important feature for classification, the classifier may confuse bright Atoll source GX 3+1 with Type-Z sources, which are observed to have high intensity values, leading to misclassification.

Furthermore, recent machine learning studies have highlighted the importance of signal-to-noise ratio (SNR) in classification tasks. Pattnaik et al. (2020) showed that lower SNR observations lead to reduced feature quality, thereby increasing classification ambiguity. In our analysis, we observe a similar trend (see Fig. 5.4), where a significant number of Pulsar sources are misclassified in the LOSO cross-validation and are associated with comparatively low SNR observations. This suggests that degradation of feature quality reduces class separability and contributes to the observed misclassifications.

However, the impact of SNR appears to be class-dependent. While Atoll sources can exhibit low-SNR observations, their classification performance remains relatively robust. This may be attributed to the presence of higher SNR observations within the Atoll class, that allows the model to learn a more well-defined representation of their feature distribution during training. In contrast, the lack of high-SNR samples for Pulsar sources may limit the model's ability to learn a clean, separable representation, making them more susceptible to misclassification under

noisy conditions. This suggests that both **data quality** and the **distribution of SNR across classes** play an important role in determining classification performance.

5.2 State Classification for Atoll sources

As mentioned in section 3.3, for state classification, we mapped the power color and RMS variability amplitude values obtained from an entire observation to the hardness ratios and normalized intensity values extracted from every 512s segment of the same observation. Because of this *many-to-one* mapping, we trained the algorithms with the data grouped observation-wise. This ensures that no bias is introduced into the classification due to segments of the same observations being split between the training and test samples with shared spectral parameters.

With RBF kernel, gamma value 0.01, and C (regularization parameter) to be 100 (refer sec 4.3.2), we used `StratifiedGroupKFold` module to achieve observation-based grouping with 80-20 train-test split and `Undersampler` to obtain consistent and equal representation of all classes among all the folds. We trained the classifier with 50 random seeds for train-test splitting and obtained class-wise accuracy scores. Fig 5.5 provides the class-wise precision, recall, and F1-score. More-

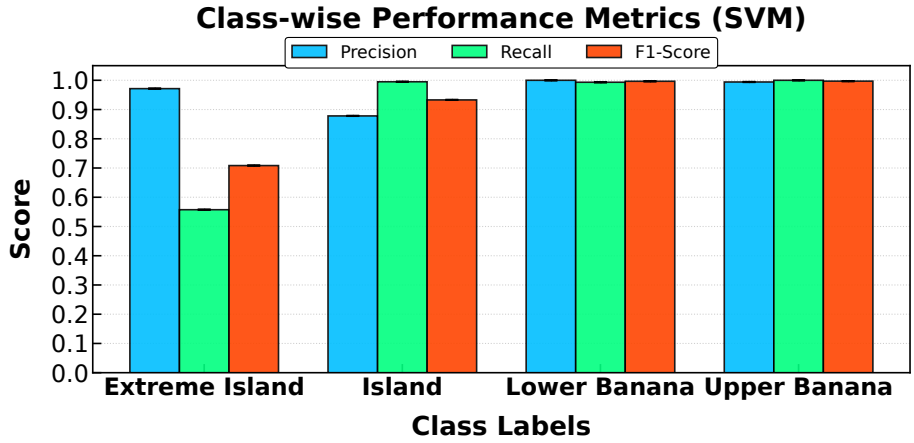


Figure 5.5: Figure shows class-wise classification performance scores (precision, recall, and F1-score, respectively) for state-classification.

over, the confusion matrix and the classifier's performance score are shown in Fig 5.6 and Table 5.3, respectively.

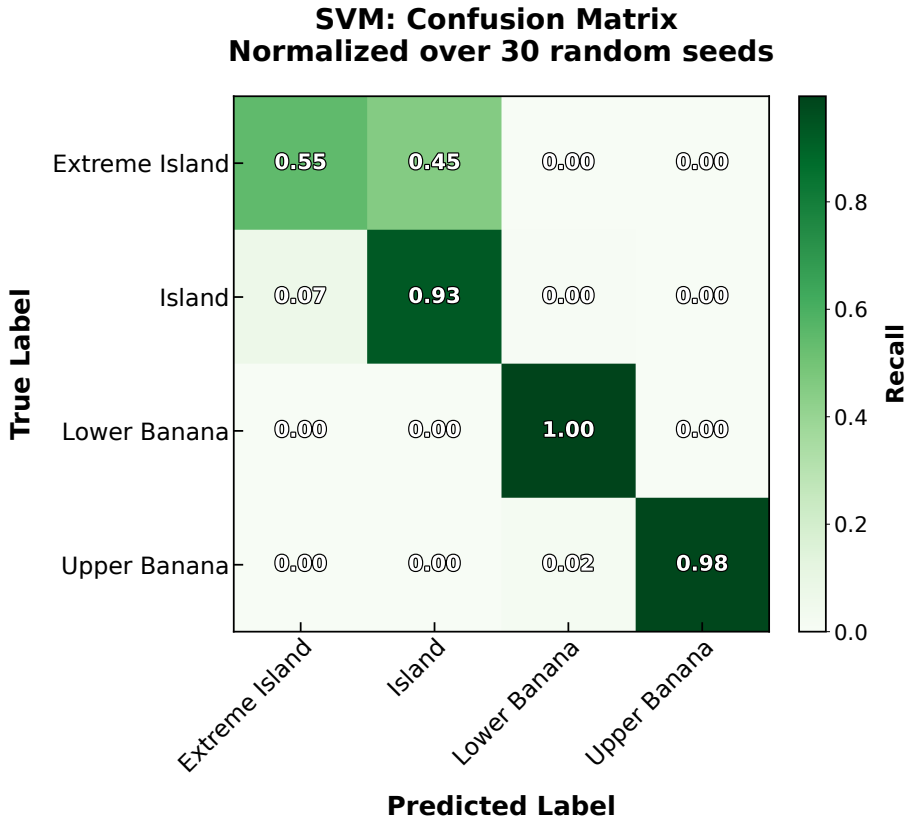


Figure 5.6: Confusion matrix for the SVM method used for state classification.

Table 5.3: Overall classification performance metrics for the state classification using the SVM model. Errors represent the standard deviation (1σ) across all cross-validation folds.

Metric	Value
Overall Accuracy	$93.6 \pm 0.6\%$
F1-score	$88.1 \pm 0.8\%$
Cohen's Kappa (κ)	0.91 ± 0.08

Based on the provided results, the Support Vector Machine performs well at classifying spectral states of the Atoll sources. From the Fig 5.5, the precision and recall values for the Extreme Island class are 1.0 and 0.55, respectively. This suggests that there are no false positives, but also that the model is failing to identify 45% of the actual candidates in this class. This shows that the model is very specific and strict in classifying data points in this class. The same can be inferred from the confusion matrix for the Extreme Island states, given that many (45%) false negatives are classified as Island states. Moreover, for the other states (Banana and Upper Banana), the model provides satisfactory

results.

We performed Leave-One-Source-Out cross-validation to evaluate whether the model can predict completely unknown observations. Fig 5.7 provides the plot of accuracy for each left-out source.

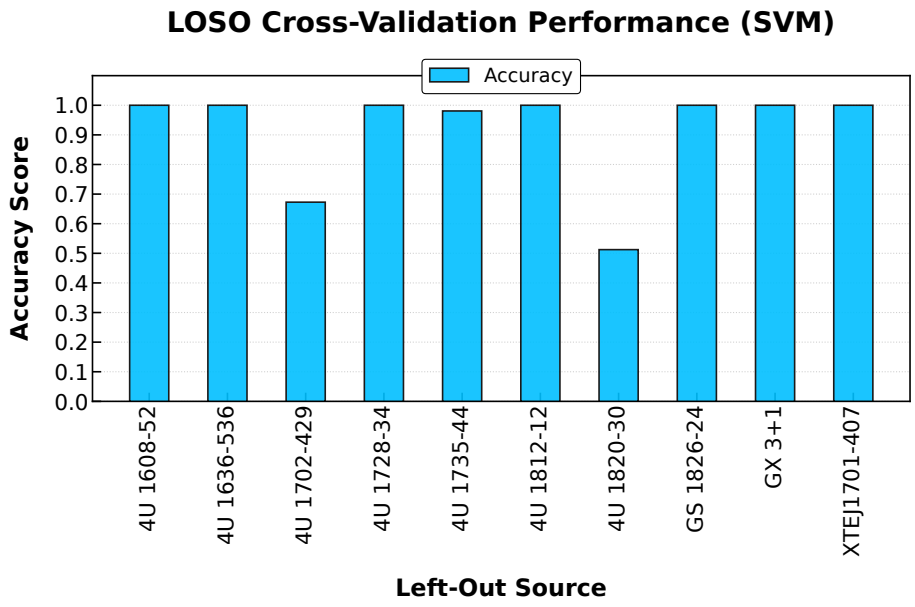


Figure 5.7: Figure shows the accuracy of state classification under the Leave-One-Source-Out method.

Overall, the classifier achieved an accuracy of 91.6 ± 5.5%. As discussed earlier, there is confusion between the Extreme Island and the Island states. A similar trend is observed in the LOSO analysis for 4U 1702-429, where the classification accuracy (67%) is lower than the overall average. Among the labeled Upper Banana sources, 4U 1820-30 constitutes approximately 30% of the total dataset. When this source is excluded for leave-one-source-out (LOSO) cross-validation, a significant portion of the dataset is removed, leading to poor representation of that class in the training dataset. This may, in turn, contribute to the relatively low classification accuracy (55%) for this source. Moreover, the results indicate that the model generalizes effectively and correctly classifies previously unseen samples into their respective states.

To further understand the model's performance and confusion between the spectral states, we discuss two key factors that are important in shaping the predictions made.

5.2.1 Feature Importance

Feature importance indicates which features strongly influence the algorithm's decision-making process. In this work, it is evaluated by calculating the drop in accuracy when that feature is omitted from both the training and test sets. Fig 5.8 provides a plot of feature importance scores vs. the features.

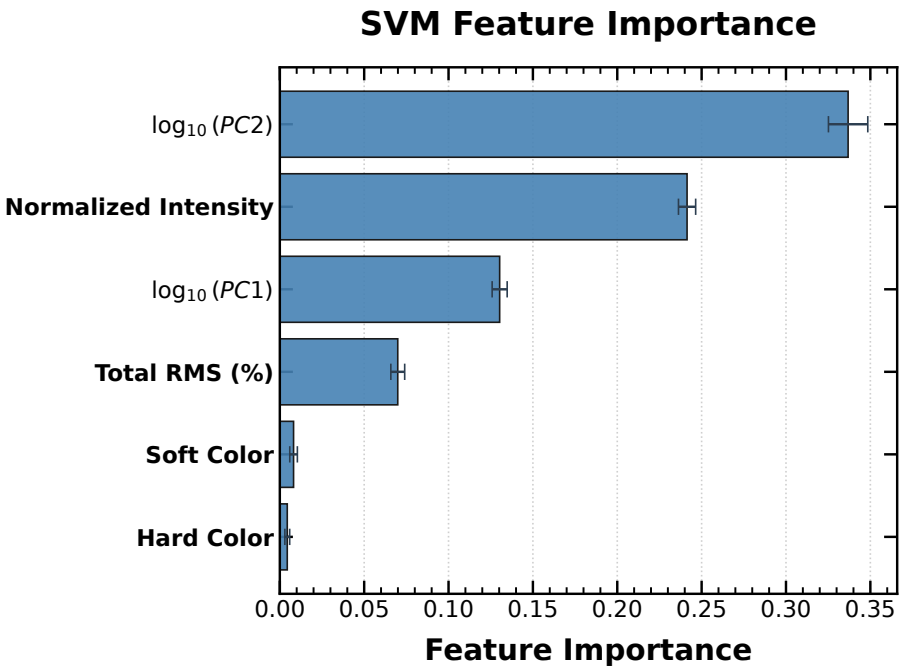


Figure 5.8: Feature importance scores for the input features, computed as the decrease in accuracy when that feature is excluded.

From the Fig. 5.8, it is evident that the power-color features and the normalized intensity emerge as the most important features in terms of their contribution to the model's performance. Spectral hardness features, which are widely used to parameterize spectral states, contribute less significantly.

It is expected that PC2 will be more dominant than PC1. This is because the way power colors are defined (see sec. 3.3), PC2 is calculated as the ratio of the variability in the frequency bands 0.031-0.25 Hz and 2-16 Hz. The 2–16 Hz frequency band captures significant state-dependent intrinsic variability, including broadband noise and quasi-periodic oscillations that become prominent in harder spectral states (van der Klis, 2006). The fractional rms variability (computed from power spectra) is highly state-dependent and increases as the source

moves from a soft state to a comparatively harder state. As a result, PC2 captures changes in the accretion dynamics across spectral states and therefore serves as an important feature for classification.

This result is consistent with previous studies, which have established that these states are better characterized by the combination of timing and spectral properties (Hasinger & van der Klis, 1989; Belloni et al., 2005). Also, power-color features have served as an effective distinguisher among states in a model-independent manner (Heil et al., 2015).

Therefore, the feature importance suggests that while spectral features provide a primary framework for state classification, the addition of timing-variability features refines the identification challenge and improves classification performance.

5.2.2 Model's behavior across state-transition regions

The evolution of the Neutron Star LMXB is a continuous process, and the source follows a specific track in the hardness-intensity diagram (HID). Although the transitions between spectral states are not sharply defined, they correspond to several regions of feature overlap throughout the evolution. To investigate how the model behaves in those regions, we analyzed multiple realizations of the train-test data set and evaluated each data point's probability of correct classification.

As from Sec . 5.2.1, intensity and the power-color features dominate the model's predictive behavior. Thus, we plotted the hardness-intensity diagram (HID) and the power color diagram (PCD), as shown in Fig 5.9, where each data point is mapped to its corresponding probability value.

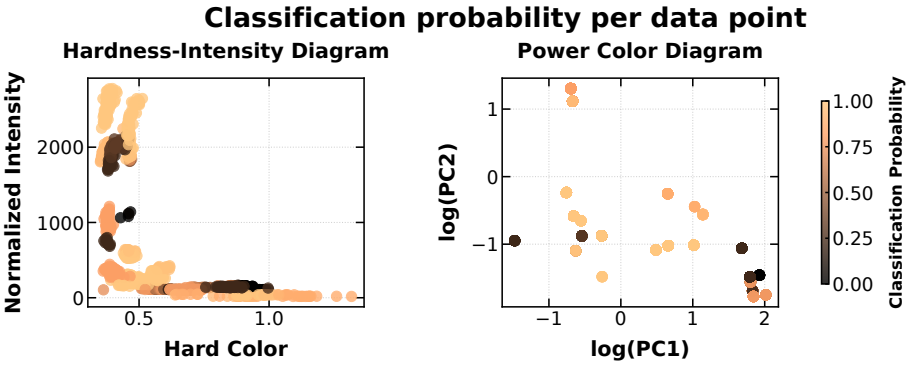


Figure 5.9: Figure provides the HID and PCD projection, here the color bar represents the classification probability for each sample.

We can observe from the figure that the model correctly identifies the well-defined states, while it also struggles in the intermediate regions: between the Upper Banana and Banana branches, and between the Extreme Island and Island branches. **This suggests that the uncertainties associated with the model are not random but are well-aligned with the source's evolution. This performance is thus consistent with the physical behavior of Atolls.**

To further assess the model's robustness, we trained the SVM on the entire labeled dataset and used it to obtain predictions for the unlabeled data. Fig 5.10 provides the predicted results for HID and PCD projection of the feature sample space.

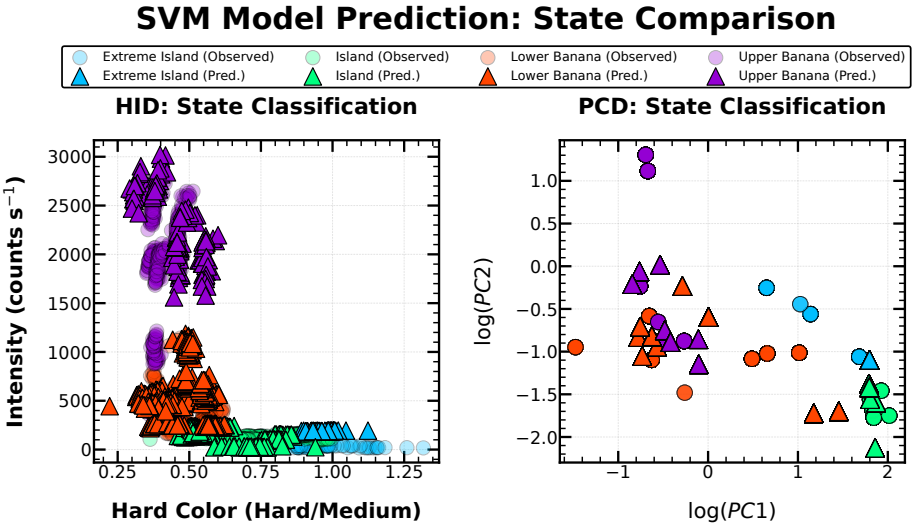


Figure 5.10: The figure shows SVM predictions on the unlabeled subset of the feature space. The SVM classifier is trained on the entire training dataset.

As shown in Fig 5.10, the predicted classification closely follows the expected evolutionary track in the HID/PCD, with the deviations also seen in the transition regions. This suggests that the model correctly captures the state evolution with good consistency.

Thus, we have examined the behavior of the best-performing algorithms for each classification type, both with labeled and unlabeled data (for state classification). In the following section, we present a comparative analysis of the models considered in this study.

5.3 Comparing Algorithms

We evaluate the performance of various algorithms, and the results of this comparison are discussed below.

5.3.1 For Source-Class classification

For source-class classification, we employed the same methodology during hyperparameter tuning and model evaluation, The training was repeated across 50 random seeds to obtain stability and robustness.

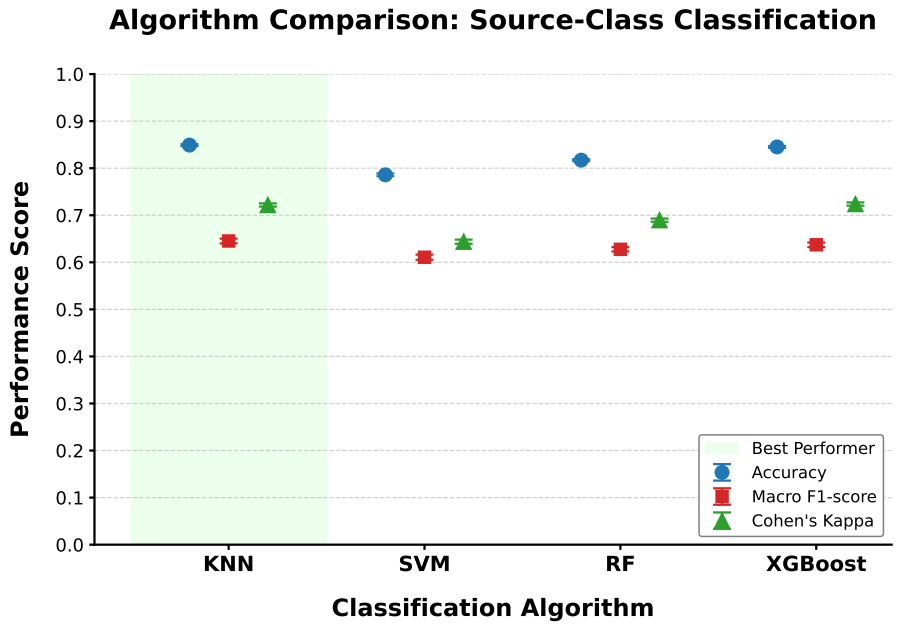


Figure 5.11: Plot of performance metric vs different models is shown for source-class-based classification. The algorithms on the x-axis are (left to right) K-nearest Neighbor, Support Vector Machine, Random Forest, and eXtreme Gradient Boosting.

Fig 5.11 shows the performance metrics vs. the chosen model. It is clear that the **K-nearest neighbors** method provides the best and most stable performance. As the distance-based KNN relies on local clustering in feature space, better performance indicates that the sources form relatively well-separated clusters when classified using spectral hardness along with intensity features.

5.3.2 For State classification

For state classification, the same evaluation was used to construct the training and test samples, and the algorithms were tested over 250 real-

izations, ensuring that each data point was evaluated at least once.

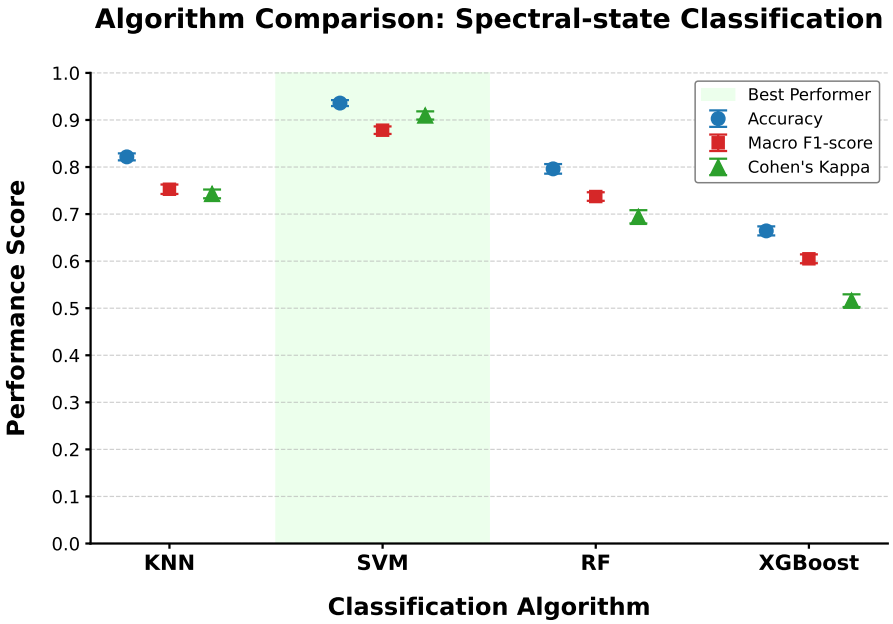


Figure 5.12: Plot of performance metric vs different models is shown for spectral state classification of atoll sources. The algorithms on the x-axis are (left to right) K-nearest Neighbor, Support Vector Machine, Random Forest, and eXtreme Gradient Boosting.

Fig 5.12 provides the comparison plot of the algorithm's performances. The best performance is achieved with an SVM using the radial basis function (RBF). This can be understood in terms of the underlying feature space, where data related to spectral states follow a continuous, nonlinear distribution. Using the kernel method, SVM can easily distinguish between classes and maintain separation for continuously evolving data. The other methods, like KNN, are highly sensitive to data points density and distribution in the feature space, which works inefficiently in this case because of overlapping data. Decision tree-based algorithms are not very effective, as classification in these methods depends heavily on splitting features along a single axis. This makes them ineffective for classifying nonlinear data.

Having discussed the results of the supervised learning procedure used for state classification, we will now present the results obtained from the unsupervised clustering algorithm GMM and discuss them in the following subsection.

5.3.3 Results from Unsupervised Learning: Gaussian Mixture Model

The Gaussian Mixture Model assigns each data point a probability of belonging to a particular cluster. From the hyperparameter tuning, GMM is optimized for 5 clusters and gives the best Adjusted Rank Index. The primary objective of using an unsupervised technique was to determine whether the feature clusters could represent the spectral states.

To identify this, we construct two overlap matrices. The first matrix provides information about how the labeled states are distributed among the clusters, and the second matrix shows how, within each cluster, the labeled data points are distributed.

The overlap matrices (ref Fig 5.13) show that the GMM clustering is closely aligned with labeled spectral states. In particular, the clusters associated with states Upper Banana and Extreme Island are largely pure, suggesting that those states occupy well-separated regions, which is consistent with their accretion pattern.

The fact that, for the intermediate states (Lower Banana and Island), the clusters are mostly dominated by labeled states, but the states are distributed across the clusters, can be explained by the unsupervised method here, which aims to obtain finer clusters. This suggests that the clustering algorithm is unable to resolve regions with limited labels. As the evolution of the source's timing and spectral properties is a continuous process, this behavior of an unsupervised learning algorithm was expected.

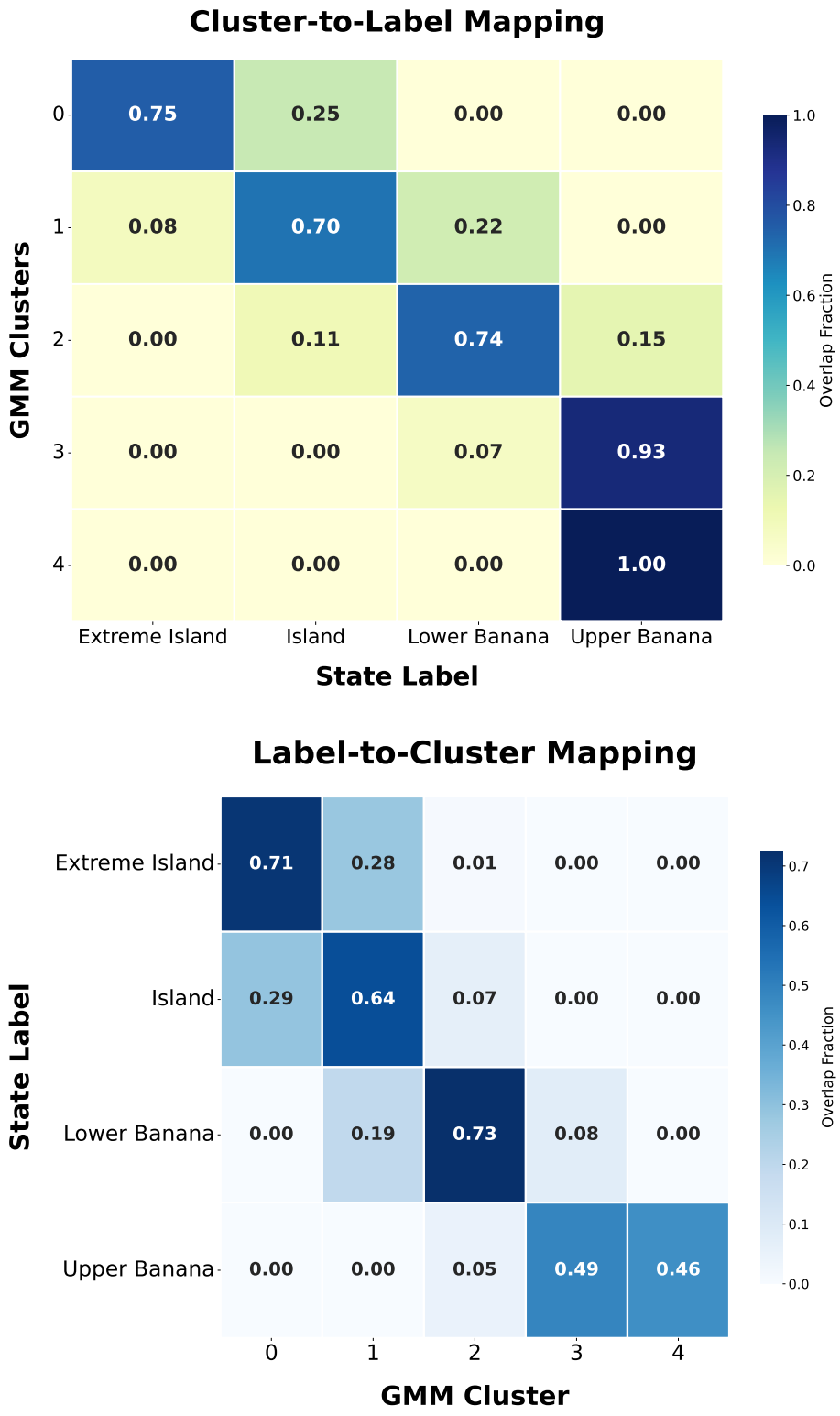


Figure 5.13: Overlap matrices showing the relationship between GMM clusters and labeled spectral states. The upper panel shows cluster-wise state composition, while the lower panel gives the distribution of each state across clusters.

Chapter 6

Conclusion

This work demonstrates that the machine learning methods can effectively classify neutron star low-mass X-ray binaries based on their spectral and variability properties. We used archival event-mode data from the LAXPC instrument onboard *AstroSat* and analyzed 35 neutron-star low-mass X-ray binaries (NS–LMXBs) across 133 observations to classify sources into their respective classes and identify spectral states within a selected source class (Atolls).

To do this, we constructed a feature sample space that combines basic spectral information, such as intensity and hardness ratios, with variability-based quantities derived from the power density spectrum, including power colors and fractional rms. These features together capture the evolution of the source across different accretion states and provide a compact representation of both spectral and timing behavior. We perform two classification tasks in this work: source-class classification, which distinguishes between Atoll, Type-Z, and Pulsar systems, and spectral-state classification, which identifies the accretion state within the Atoll source class.

Source Class Classification:

- The machine learning models are able to classify neutron star LMXBs into their respective source classes with good overall accuracy, indicating that the chosen feature space captures the primary differences between Atoll, Type-Z, and Pulsar systems using the features: hard color, soft color, and normalized intensities.
- Among the tested algorithms, K-nearest neighbor (KNN) achieves the highest classification accuracy score of $84.9 \pm 0.2\%$, which

suggests that local clustering in feature space is important for distinguishing between source classes.

- The classification performance is not uniform across all classes. Higher misclassification is observed for Pulsar sources and the Atoll + Z category. This can be attributed to intrinsic similarities in their spectral properties, which lead to overlapping regions in the feature space and limit the class separability.

- Feature importance analysis shows that normalized intensity is the most dominant feature for classification. This dominance may bias the classifier; for example, the bright Atoll source GX 3+1 is misidentified as a Type-Z source, possibly because of its similar high-intensity values.

- The results indicate that data quality significantly impacts classification performance. Pulsar sources, which are typically associated with low-SNR observations, show a higher misclassification rate due to degraded feature quality and reduced class separability. In contrast, Atoll sources maintain robustness at low SNR, probably because higher-SNR observations within the class help the model learn a clearer representation. The lack of high-SNR Pulsar samples makes them more prone to misclassification in noisy conditions.

State Classification for Atolls:

- Among the models tested, SVM with a radial basis function (RBF) kernel performs the best with an accuracy score of $93.6 \pm 0.6\%$. This suggests that the SVM can easily distinguish between classes and maintain separation even when data evolve continuously. SVM with the RBF kernel is well-suited to handling nonlinear feature spaces, which likely contributes to its strong performance on this dataset.

- For this analysis, timing-variability features such as power colors and fractional rms were included in the feature dataset. These variability features play a dominant role in distinguishing between spectral states, outperforming traditional spectral hardness features in terms of feature importance.

- The classification performance is robust for well-defined states such as Banana and Upper Banana, while higher confusion is observed between Extreme Island and Island states. This confusion can be attributed to the continuous evolution of the source, leading to overlap in the feature space.

- The model's behavior across the HID and PCD indicates that classification uncertainties are not random but are aligned with the physical state-transition regions. The confusion arises primarily in these transitional regions, where feature overlap is expected due to the continuous nature of state evolution.

- In addition to supervised classification, the unsupervised Gaussian Mixture Model (GMM) analysis identifies five clusters in the feature space that largely map to distinct spectral states. One of these clusters appears to lie in the intermediate region and may represent transitional behavior between overlapping states. This suggests that, even without labels, the data in the feature space naturally organizes in a manner consistent with the source's evolution during the outburst.

This analysis is based on a sample drawn from the *AstroSat* data archive. The sample size is smaller compared to large datasets from missions such as *RXTE*, which provide richer temporal coverage and longer continuous monitoring of sources. Despite this, the dataset is sufficient to capture the key spectral and variability characteristics of neutron star LMXBs. The results show that machine learning methods can reliably identify the underlying behavior and evolutionary patterns of these systems even with this dataset.

Chapter 7

Future directions

Based on the methods and results in this work, we propose the following as future plans to improve classification results.

- So far, we have used multiple machine learning techniques for source class classification and for the state classification of the Atoll sources. More advanced models, such as neural networks and transformer-based architectures, can be employed to further improve classification performance. The overlapping regions can be better resolved by these approaches, thereby increasing robustness across the different sources.

- For the Type-Z sources, the primary challenge is to obtain reliable timing-variability features due to their highly non-stationary nature. The construction of the power density spectrum requires the assumption of statistical stationarity within each light curve segment. Therefore, a dedicated analysis at the individual-observation level is necessary. In particular, careful segmentation of the light curve into quasi-stationary intervals is essential to ensure the validity of the PDS estimation. Such an approach improves the robustness of the extracted timing features and can further aid in the classification of Type-Z spectral states.

- The consistency between the supervised labels for Atoll state classification and the intrinsic data can be explored using principal component analysis (PCA) on segment-wise power density spectrum (PDS) features. The segments can be projected into a reduced-dimensional feature space and compared with the classifier predictions to check whether the learned decision boundaries

are consistent with the distribution of features in the PCA space.

- Recent studies highlight that Bayesian approaches to power spectral estimation can be more reliable than traditional Fourier-based periodogram methods (Hübner et al., 2022). In this context, using Bayesian approaches, power density spectrum estimation can be used to extract more robust timing-variability features for classification purposes.

- Additionally, Black-hole LMXBs data can be incorporated in the sample space to enable broader classification. Furthermore, the spectral states of BH LMXBs can also be classified using ML models.

Bibliography

Agrawal, V. K., Nandi, A., & Katoch, T. 2022, arXiv e-prints, doi: 10.48550/arXiv.2209.08275

Armas Padilla, M., Rodríguez-Gil, P., Muñoz-Darias, T., et al. 2022, The Astrophysical Journal Letters, 931, L9, doi: 10.3847/2041-8213/ac6d54

Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Mantelet, G., & Andrae, R. 2018, The Astronomical Journal, 156, 58, doi: 10.3847/1538-3881/aacb21

Ballantyne, D. R., & Strohmayer, T. E. 2004, The Astrophysical Journal, 602, L105, doi: 10.1086/382703

Balman, S. 2010, Monthly Notices of the Royal Astronomical Society, 407, 1895, doi: 10.1111/j.1365-2966.2010.17033.x

Belloni, T., Homan, J., Casella, P., et al. 2005, Astronomy and Astrophysics, 440, 207, doi: 10.1051/0004-6361:20042457

Beri, A., Paul, B., Yadav, J. S., et al. 2018, arXiv e-prints, doi: 10.48550/arXiv.1811.10030

Bhattacharya, S., & Bhattacharyya, S. 2025, What causes long outbursts of neutron star low-mass X-ray binaries? https://arxiv.org/abs/2504.07621

Bhattacharyya, S. 2007, Monthly Notices of the Royal Astronomical Society, 377, 198, doi: 10.1111/j.1365-2966.2007.11587.x

Bhattacharyya, S., Yadav, J. S., Sridhar, N., et al. 2018, arXiv e-prints, doi: 10.48550/arXiv.1805.05393

Bhattacherjee, S., Nath, A., Sarkar, B., et al. 2024a, The Astrophysical Journal, 971, 154, doi: 10.3847/1538-4357/ad583d

Bhattacherjee, S. S., Nath, A. A., Sarkar, B. B., Beri, A. A., et al. 2024b, arXiv e-prints, doi: 10.48550/arXiv.2406.10666

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

Campana, S., Colpi, M., Mereghetti, S., Stella, L., & Tavani, M. 1998, The Astronomy and Astrophysics Review, 8, 279, doi: 10.1007/s001590050012

Cerda-Duran, P., & Elias-Rosa, N. 2018, Neutron Stars Formation and Core Collapse Supernovae (Springer International Publishing), 1–56, doi: 10.1007/978-3-319-97616-7_1

Chattopadhyay, S., Mandal, S., & Misra, R. 2025, arXiv e-prints, doi: 10.48550/arXiv.2505.00676

Chaty, S. 2022, in Accreting Binaries, 2514-3433 (IOP Publishing), 5–1 to 5–46, doi: 10.1088/2514-3433/ac595fch5

Chen, Y.-P., Li, J., Xiong, S.-L., et al. 2022, The Astrophysical Journal, 935, 10, doi: 10.3847/1538-4357/ac7ff8

Clark, G. W. 2018, The Astrophysical Journal, 852, 121, doi: 10.3847/1538-4357/aaa1f0

Cocchi, M., Bazzano, A., Natalucci, L., et al. 2000, Observations of Eddington-limited type-I X-ray bursts from 4U 1812-12, arXiv, doi: 10.48550/arXiv.astro-ph/0003151

Cover, T., & Hart, P. 1967, IEEE Transactions on Information Theory, 13, 21, doi: 10.1109/TIT.1967.1053964

De Falco, V., Kuiper, L., Bozzo, E., et al. 2017, Astronomy & Astrophysics, 599, A88, doi: 10.1051/0004-6361/201629575

Degenaar, N., Medin, Z., Cumming, A., et al. 2014, The Astrophysical Journal, 791, 47, doi: 10.1088/0004-637x/791/1/47

Dempster, A. P., Laird, N. M., & Rubin, D. B. 2018, Journal of the Royal Statistical Society: Series B (Methodological), 39, 1, doi: 10.1111/j.2517-6161.1977.tb01600.x

Done, C., Gierliński, M., & Kubota, A. 2007, The Astronomy and Astrophysics Review, 15, 1–66, doi: 10.1007/s00159-007-0006-1

Falanga, M., Cumming, A., Bozzo, E., & Chenevez, J. 2009, Astronomy & Astrophysics, 496, 333, doi: 10.1051/0004-6361/200811091

Fender, R. P., & Hendry, M. A. 2000, Monthly Notices of the Royal Astronomical Society, 317, 1, doi: 10.1046/j.1365-8711.2000.03443.x

Frank, J., King, A., & Raine, D. J. 2002, Accretion Power in Astrophysics: Third Edition

Fryer, C. L. 1999, The Astrophysical Journal, 522, 413, doi: 10.1086/307647

Galloway, D. K., Muno, M. P., Hartman, J. M., Psaltis, D., & Chakrabarty, D. 2008, The Astrophysical Journal Supplement Series, 179, 360, doi: 10.1086/592044

Galloway, D. K., Psaltis, D., Muno, M. P., & Chakrabarty, D. 2006, The Astrophysical Journal, 639, 1033, doi: 10.1086/499579

Gasealahwe, K. V. S., Monageng, I. M., Fender, R. P., et al. 2023, Monthly Notices of the Royal Astronomical Society, 521, 2806, doi: 10.1093/mnras/stad649

Gladstone, J., Done, C., & Gierlinski, M. 2007a, Monthly Notices of the Royal Astronomical Society, 378, 13–22, doi: 10.1111/j.1365-2966.2007.11675.x

—. 2007b, Monthly Notices of the Royal Astronomical Society, 378, 13, doi: 10.1111/j.1365-2966.2007.11675.x

Gopalan, G., Vrtilek, S. D., & Bornn, L. 2015, The Astrophysical Journal, 809, 40, doi: 10.1088/0004-637X/809/1/40

Grimm, H.-J., Gilfanov, M., & Sunyaev, R. 2002, Astronomy & Astrophysics, 391, 923, doi: 10.1051/0004-6361:20020826

Grinberg, V., Pottschmidt, K., Böck, M., et al. 2014, Astronomy & Astrophysics, 565, A1, doi: 10.1051/0004-6361/201322969

Güngör, C., Güver, T., & Ekşi, K. Y. 2014, Monthly Notices of the Royal Astronomical Society, 439, 2717, doi: 10.1093/mnras/stu128

Güver, T., Boztepe, T., Ballantyne, D. R., et al. 2022, Monthly Notices of the Royal Astronomical Society, 510, 1577, doi: 10.1093/mnras/stab3422

Hasinger, G., & Klis, M. 1989, Astronomy and Astrophysics, 225, 79

Hasinger, G., & van der Klis, M. 1989, Astronomy and Astrophysics , 225, 79

Hasinger, G., van der Klis, M., Ebisawa, K., Dotani, T., & Mitsuda, K. 1990, Astronomy and Astrophysics , 235, 131

Heger, A., Cumming, A., Galloway, D. K., & Woosley, S. E. 2007, The Astrophysical Journal, 671, L141, doi: 10.1086/525522

Heil, L. M., Uttley, P., & Klein-Wolt, M. 2015, Monthly Notices of the Royal Astronomical Society, 448, 3339, doi: 10.1093/mnras/stv191

Heinz, S., Sell, P., Fender, R. P., et al. 2013, The Astrophysical Journal, 779, 171, doi: 10.1088/0004-637X/779/2/171

Hills, J. G. 1980, The Astrophysical Journal , 235, 986, doi: 10.1086/157703

Homan, J., van der Klis, M., Fridriksson, J. K., et al. 2010a, The Astrophysical Journal, 719, 201–212, doi: 10.1088/0004-637x/719/1/201

Homan, J., Van Der Klis, M., Fridriksson, J. K., et al. 2010b, The Astrophysical Journal, 719, 201, doi: 10.1088/0004-637X/719/1/201

Huppenkothen, D., Heil, L. M., Hogg, D. W., & Mueller, A. 2016, Monthly Notices of the Royal Astronomical Society, 466, 2364, doi: 10.1093/mnras/stw3190

Hynes, R. I., & Jones, E. D. 2009, The Astrophysical Journal, 697, L14, doi: 10.1088/0004-637X/697/1/L14

Hübner, M., Huppenkothen, D., Lasky, P. D., & Inglis, A. R. 2022, The Astrophysical Journal Supplement Series, 259, 32, doi: 10.3847/1538-4365/ac49ec

Iaria, R., Di Salvo, T., Del Santo, M., et al. 2016, Astronomy & Astrophysics, 596, A21, doi: 10.1051/0004-6361/201628210

In 'T Zand, J. J. M., Kuulkers, E., Verbunt, F., Heise, J., & Cornelisse, R. 2003, Astronomy & Astrophysics, 411, L487, doi: 10.1051/0004-6361:20031586

Jonker, P. G., Klis, M. v. d., Kouveliotou, C., et al. 2003, Monthly Notices of the Royal Astronomical Society, 346, 684, doi: 10.1046/j.1365-2966.2003.07123.x

Juris, N. J., James, M., & Devasia, J. 2025, Research in Astronomy and Astrophysics, 25, 055002, doi: 10.1088/1674-4527/adc789

Kashyap, U., Maccarone, T. J., Russell, T. D., et al. 2025, The Astrophysical Journal, 985, 245, doi: 10.3847/1538-4357/adceb9

Kashyap, U., Ram, B., Güver, T., & Chakraborty, M. 2021, arXiv e-prints, doi: 10.48550/arXiv.2109.14631

Keek, L., Iwakiri, W., Serino, M., et al. 2017, The Astrophysical Journal, 836, 111, doi: 10.3847/1538-4357/836/1/111

Kiel, P. D., & Hurley, J. R. 2006, , 369, 1152, doi: 10.1111/j.1365-2966.2006.10400.x

King, A. R. 1998, MNR AS

Klein-Wolt, M., & Van Der Klis, M. 2008, The Astrophysical Journal, 675, 1407, doi: 10.1086/525843

Landis, J. R., & Koch, G. G. 1977, Biometrics, 33, 159, doi: 10.2307/2529310

Lasota, J.-P. 2001, New Astronomy Reviews, 45, 449–508, doi: 10.1016/s1387-6473(01)00112-9

Lavanya, S., Thomas, N. T., Gudennavar, S. B., & Bubbly, S. G. 2024a, Monthly Notices of the Royal Astronomical Society, 534, 2783, doi: 10.1093/mnras/stae2269

Lavanya, S., et al. 2024b, Monthly Notices of the Royal Astronomical Society, 534, 2783, doi: 10.1093/mnras/stae2269

Lewin, W. H. G., van Paradijs, J., & van den Heuvel, E. P. J. 1995, X-ray Binaries (Cambridge: Cambridge University Press)

Lu, Y., Li, Z., Yu, W., Pan, Y., & Falanga, M. 2024, The Astrophysical Journal, 969, 15, doi: 10.3847/1538-4357/ad4d86

Ludlam, R. M., Shishkovsky, L., Bult, P. M., et al. 2019, The Astrophysical Journal, 883, 39, doi: 10.3847/1538-4357/ab3806

Maitra, D., & Bailyn, C. D. 2004, The Astrophysical Journal, 608, 444–453, doi: 10.1086/382746

Markwardt, C. B., Miller, J. M., & Wijnands, R. 2002, The Astronomer's Telegram, 110, 1

Meyer, F., & Meyer-Hofmeister, E. 1981, Astronomy and Astrophysics, 104, L10

Migliari, S., & Fender, R. P. 2006, Monthly Notices of the Royal Astronomical Society, 366, 79, doi: 10.1111/j.1365-2966.2005.09777.x

Migliari, S., Miller-Jones, J. C. A., Fender, R. P., et al. 2007, The Astrophysical Journal, 671, 706–712, doi: 10.1086/522516

Mukherjee, D., Bult, P., van der Klis, M., & Bhattacharya, D. 2015, Monthly Notices of the Royal Astronomical Society, 452, 3994, doi: 10.1093/mnras/stv1542

Orosz, J. A., & Kuulkers, E. 1999, Monthly Notices of the Royal Astronomical Society, 305, 132, doi: 10.1046/j.1365-8711.1999.t01-1-02420.x

Paczyński, B. 1971, , 9, 183, doi: 10.1146/annurev.aa.09.090171.001151

Park, T., Kashyap, V. L., Siemiginowska, A., et al. 2006, The Astrophysical Journal, 652, 610, doi: 10.1086/507406

Parmar, A. N., Gottwald, M., Van Der Klis, M., & Van Paradijs, J. 1989, The Astrophysical Journal, 338, 1024, doi: 10.1086/167255

Pattnaik, R., Sharma, K., Alabarta, K., et al. 2020, Monthly Notices of the Royal Astronomical Society, doi: 10.1093/mnras/staa3899

Pradhan, P., Paul, B., Bozzo, E., Maitra, C., & Paul, P. C. 2023, arXiv e-prints, doi: 10.48550/arXiv.2307.16566

Rankin, J., La Monaca, F., Di Marco, A., et al. 2024, The Astrophysical Journal Letters, 961, L8, doi: 10.3847/2041-8213/ad1832

Remillard, R. A., & McClintock, J. E. 2006, Annual Review of Astronomy and Astrophysics, 44, 49, doi: 10.1146/annurev.astro.44.051905.092532

Ricketts, B. J., Steiner, J. F., Garraffo, C., Remillard, R. A., & Huppenkothen, D. 2023, Monthly Notices of the Royal Astronomical Society, 523, 1946–1966, doi: 10.1093/mnras/stad1332

Roy, P., Beri, A., & Mondal, A. S. 2022, arXiv e-prints, doi: 10.48550/arXiv.2202.07379

Salgundi, A., Bala, S., Raman, G., Pathak, U., & Bhalerao, V. 2024, "Bursts, Beats, and Beyond": Uncovering the landscape from accretion to ignition of 4U 1728–34 using *AstroSat*, doi: 10.48550/arXiv.2412.06644

Sanna, A., Ferrigno, C., Ray, P. S., et al. 2018, Astronomy & Astrophysics, 617, L8, doi: 10.1051/0004-6361/201834160

Shakura, N. I., & Sunyaev, R. A. 1973, Astronomy and Astrophysics , 24, 337

Shapiro, S. L., Lightman, A. P., & Eardley, D. M. 1976, The Astrophysical Journal , 204, 187, doi: 10.1086/154162

Sharma, R., Jaleel, A., Jain, C., et al. 2018, Monthly Notices of the Royal Astronomical Society, 481, 5560, doi: 10.1093/mnras/sty2678

Sreehari, H., & Nandi, A. 2021, Monthly Notices of the Royal Astronomical Society, 502, 1334, doi: 10.1093/mnras/stab151

Straaten, S. v., Klis, M. v. d., & Wijnands, R. 2005, The Astrophysical Journal, 619, 455, doi: 10.1086/426183

Sunyaev, R. A., & Titarchuk, L. G. 1980, Astronomy and Astrophysics , 86, 121

Thomas, N. T., Gudennavar, S. B., & Bubbly, S. G. 2023a, Monthly Notices of the Royal Astronomical Society, 521, 433, doi: 10.1093/mnras/stad555

Thomas, N. T., Lavanya, S., Gudennavar, S. B., & Bubbly, S. G. 2025, Journal of High Energy Astrophysics, 45, 250, doi: 10.1016/j.jheap.2024.12.007

Thomas, N. T., et al. 2023b, Monthly Notices of the Royal Astronomical Society, 521, 433, doi: 10.1093/mnras/stad555

van der Klis, M. 2006, Rapid X-ray Variability, ed. W. H. G. Lewin & M. van der Klis (Cambridge University Press), 39–112

Vapnik, V. N. 1995, The Nature of Statistical Learning Theory (New York: Springer), doi: 10.1007/978-1-4757-2440-0

Varun, M., et al. 2024, Monthly Notices of the Royal Astronomical Society, 529, 2234, doi: 10.1093/mnras/stae636

Vrtilek, S. D., & Boroson, B. S. 2013, Monthly Notices of the Royal Astronomical Society, 428, 3693, doi: 10.1093/mnras/sts312

Wang, Z., & Chakrabarty, D. 2004, The Astrophysical Journal, 616, L139, doi: 10.1086/426787

Yadav, J. S., Agrawal, P. C., Antia, H. M., et al. 2016, in , Edinburgh, United Kingdom, 99051D, doi: 10.1117/12.2231857

Yan, Z., & Yu, W. 2015, The Astrophysical Journal, 805, 87, doi: 10.1088/0004-637X/805/2/87

Zand, J. J. M. i. t., Bassa, C. G., Jonker, P. G., et al. 2008, Astronomy & Astrophysics, 485, 183, doi: 10.1051/0004-6361:200809361