JOURNAL OF
EMERGING INVESTIGATORS

Article

Analysis of quantitative classification and properties of X-ray binary systems

Devan Kuppusamy[1], Praveen Rajkumar[1]*, Chris Garcia[1]*, Shyamal Mitra[2]

[1] Westlake High School, Austin, Texas

[2] College of Natural Sciences, University of Texas at Austin, Austin, Texas

* These authors contributed equally to this work

SUMMARY

X-ray astronomy is a field of research focusing on astronomical objects that emit X-ray radiation. The first interstellar X-ray source, Scorpius X-1, was discovered in 1962. Since then, hundreds of X-ray binaries (XRBs) have been discovered inside the Milky Way, with many unusual properties. These binary systems are currently classified as either high- or low-mass XRBs. However, it is useful to the field to have a classification system consisting of more properties than just mass to have a more descriptive classification of a stellar object. We aimed to analyze the patterns in several variables contributing to an XRB's properties to find better-fit ways to classify these systems. We hypothesized that there would be a large difference in classes of XRBs across several variables, not just mass, but including the characteristics of a companion star. Utilizing several machine learning algorithms, we analyzed several variables describing an XRB. Dimensionality reduction was used to find correlations between the most important variables defining a star such as temperature, orbital period, metallicity, and redshift, and the classification groups each star falls under. We found that temperature is a quantitative parameter to classify these systems. There are three main groups of temperatures that XRBs fall into. These being stars with a metallicity <-2 and <100000K, those with a metallicity >-2 and <100000K, and those with a metallicity >-2 and >100000K. Characterizing XRBs are important to find star parameters that explain the chemical evolution of the galaxy and understanding of how binary parameters evolve toward detectable mergers.

INTRODUCTION

The study of X-ray binaries (XRBs) is of great interest to the field of astrophysics due to the information they provide on the evolution of high-mass companion stars and the properties of the accretion disks (disks of matter orbiting a high mass object) that surround them (1). XRBs are systems consisting of two stars that orbit around one center of mass and emit X-ray wavelength radiation. These systems emit almost all of their radiation in the form of X-ray radiation, with minimal radiation emitted in the visible wavelength, with XRBs being the foremost source of X-rays in the Milky Way (2). XRBs are binary systems comprising one compact object – such as black holes, neutron stars, or (unconventionally) white dwarfs – and are also accompanied by a main sequence star, with the sun being an example of such (1). The emission of X-rays is due to accretion disks formed from matter being pulled from one stellar object to the other and heating up as it orbits (1). Understanding the variety of stellar objects that comprise XRBs gives a fuller picture of the conditions necessary for accretion disks to form.

Due to how hot the matter that orbits a star becomes (the accretion disk), it emits X-rays. This means these are some of the most extreme objects in the universe due to the extreme temperature. The companion objects, stars from which the compact objects draw matter, have been observed to be of a wide range of different classifications, such as O or B stars (10,000 K (Kelvin) – 25,000 K), Wolf-Rayet Stars (stars at an endpoint of stellar evolution before supernova), white dwarfs (incredibly dense stars where nuclear fusion has ceased), and red-giant stars (stars of later stellar evolution under 5,000K with inflated radius) (3). Currently, XRBs are classified as either high-mass X-ray binaries (HMXB) (greater than 10 solar masses) or low-mass X-ray binaries (LMXB) (less than one solar mass); however, a lesser-used classification of intermediate X-ray binaries (IMXB) does exist (1). It is important to note that historically, XRBs have been classified based on the mass of their companion stars (2).

Researching the conditions needed to enable accretion disks to form allows for further examination into active galactic nuclei (AGN), a region in the center of a galaxy that emits a large amount of energy, typically black holes, and a more specific investigation into their relativistic jets, which are ionized matter emitted from an AGN near the speed of light (4). Both AGNs and XRBs have accretion disks. Using machine learning to examine relationships between many factors in forming XRBs, such as effective temperature, metallicity, orbital period, and redshift, can help investigate relationships between these factors, apart from the mass, to classify these systems (5). XRBs offer insight into the final point in a binary star system's evolution and demonstrate how a star exists under the extreme conditions for accretion disk formation, as matter is heated to billions of degrees Kelvin, especially with a compact object such as a black hole. However, analyzing the patterns in astronomical data such as this is challenging for humans due to the high dimensional data sets utilized, supporting a machine learning approach (4).

The number of variables required to analyze celestial body data is well-suited for machine learning (6). In this study, we utilize both supervised and unsupervised machine learning. Unsupervised machine learning is ideal for clustering algorithms, which we utilize to uncover clusters in data by two

https://doi.org/10.59720/24-090

or more variables indicating patterns in data other than mass difference (5). Supervised machine learning was also used by pulling from existing XRB databases to decipher whether XRBs can be classified into current classification systems based on factors apart from mass. This information was then plotted on an Aitoff projection (2-D projection of the sky) to determine whether there is a correlation between the standardized median value and a star's deviation from that median across several variables **(Figure 1)**.

In our study, we essayed to understand whether it would be possible to classify XRBs on quantitative factors other than mass easily. Utilizing many factors like temperature and metallicity to describe an XRB would be more constructive and practical for representing the true nature of an XRB rather than just mass (6). We also wanted to determine if we could predict the mass of XRBs based on these other factors, and whether the combination of the mass of XRBs with these other factors could classify and describe these objects better. We hypothesized that there would be a notable difference in the types of XRBs across several factors, particularly with a considerable weight placed on the properties of the system's companion star. We found three major temperature groups of XRBs that can aid in creating a classification system. Low-mass stars predominantly have temperatures between 5,000–10,000 Kelvin, while HMXBs have temperatures greater than 10,000 Kelvin. The temperature had the most significant influence on successfully classifying an XRB. This result showed that the most critical determinant of the type of XRB was not metallicity. However, there was significantly more variability in the temperature of HMXBs with metallicities above -1 than LMXBs across all metallicity ranges. Consequently, we could find distinct groups of XRBs based on temperature, metallicity, and orbital period. Temperature was the most indicative, followed by metallicity, then orbital period. It is possible to quantitatively classify XRBs on factors other than mass, meaning that a more detailed classification can be used to describe an XRB with variables besides mass alone.

RESULTS

We started with an open-source dataset called XRBCats which consists of 517 XRBs, with 169 HMXBs and 348 LMXBs, with each source having right ascension, declination, mean distance (parsecs), pulsation period, orbital period, effective temperature, and mass as factors. We then added metallicity and redshift as factors, pulling data from the Gaia Archive, a catalog containing information on stellar objects. Metallicity measures the abundance of all elements heavier than Helium present in a star. Generally, higher metallicity stars are younger, brighter, and higher in temperature (4). Redshift is how much the wavelength of light is stretched as it moves away from the Earth and can tell the velocity a star moves away from the Earth and its distance (1). We plotted the right ascension and declination on an Aitoff projection to visualize the distribution of the HMXB vs LMXB objects and determine if they were Pop or Pop II systems **(Figure 1)** (7). Pop I systems are younger, blue, more metal-rich, and typically located in the plane of the galaxy. Pop II systems are older, redder, less metal-rich, and located either in the halo or bulge of the galaxy. This dataset followed patterns previous researchers presented and was determined to be suitable for our experiments (7).

t-SNE comparison

t-SNE (t-distributed stochastic neighbor embedding) is an unsupervised machine-learning dimension reduction procedure that allows the user to visualize high-dimensional data in two-dimensional space while retaining local properties (8). As we dealt with many variables in this experiment, we needed to be able to visualize relationships between the many factors. We identified effective temperature, metallicity, orbital period, and type of XRB as applicable factors for comparison. We found temperature to be effective as it is a widely used star comparison metric; metallicity as it determines the rate of stellar evolution; orbital period as it

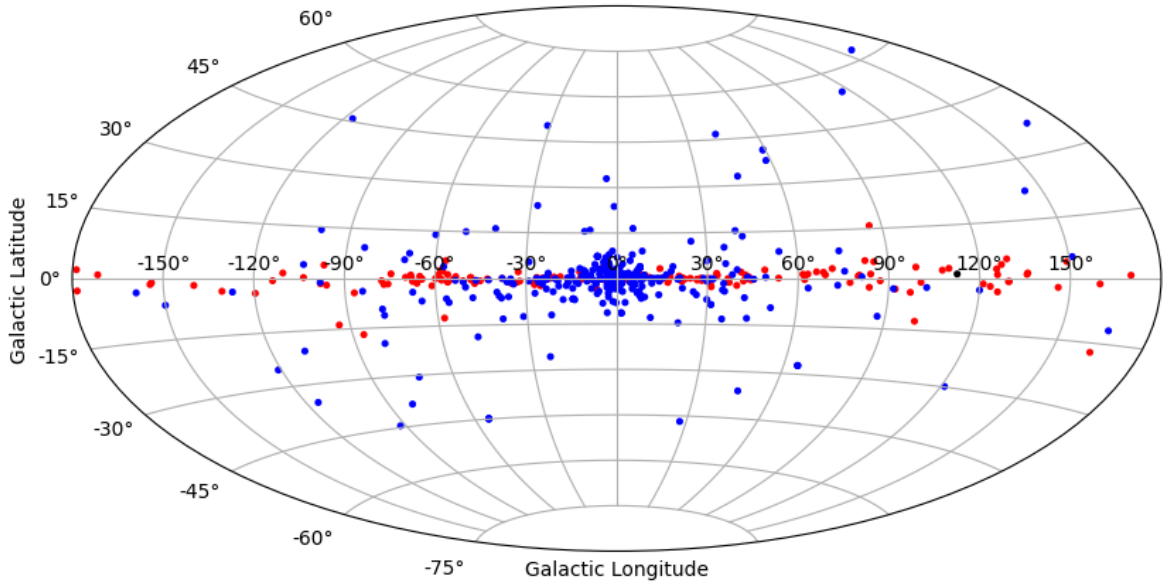

Figure 1. Aitoff projection of X-ray binaries. Physical Location of high-mass (red) and low-mass (blue) X-Ray binaries projected onto an Aitoff map projection. Right ascension and declination were converted to galactic longitude and latitude, respectively, and plotted on a map; high-mass X-Ray binaries stay within the galactic plane while low-mass X-Ray binaries spread further out.

https://doi.org/10.59720/24-090

gives insight into the relative masses of the binary system; and XRB type to determine the effectiveness of our classifier. This procedure reduced the number of data points used from 517 to 110. We found that when utilizing the t-SNE algorithm, there are three main data points as compared by either HMXB or LMXB, as shown **(Figure 2)**. One of the main groups is a set of overlapping points. A positive linear relationship exists between t-SNE component 1 and t-SNE component 2. Examining the importance of this feature, we found that effective temperature had the most significant contribution (-0.219), followed by metallicity at -0.17. From this observation, we overlaid the standardized temperature data over the current points. We found that as the t-SNE 1 and 2 values become more extreme, the temperature deviates farther from the median. Our study revealed three main groups of data **(Figure 3).**

Temperature and metallicity comparison

After examining the relationship shown in the t-SNE comparison, we decided to utilize k-means clustering due to its simplicity, ease of use, and the characteristics of this data set having few outliers, which is suitable for k-means (9). We found three main temperature groups: the first was from about 2000K - 10000K, the second from 10000K - 25000K, and the third from 25000K - 45000K. As there was also a significant dependency on metallicity in the initial t-SNE comparison, we decided to compare the effective temperature to metallicity. We found an independent section of HMXBs that generally only have metallicities greater than -1. This finding contrasts LMXB, whose metallicities range from -4 to 1, but effective temperatures do not rise above 10000K. There was an overlap, as seen in the t-SNE plot where these two groups intersect **(Figure 4)**.

Mass prediction and current classification

To evaluate the current classification system of HMXBs and LMXBs, we decided to try and predict the mass of a system using a variety of factors such as metallicity, temperature, mass-related factor, and orbital period, but not discrete mass values (10). We found that a logarithmic classifier was the best fit for predicting the mass of the system **(Figure 5).** The residual graph shows that most predicted masses were near a 0 solar mass difference. Residual is calculated as actual value minus predicted value. This demonstrates high accuracy in the classifier, as most estimations are near the actual star mass. Using mass-related variables, we could predict a star's mass with about 86% accuracy. For example, our model identified low-mass stars accurately but struggled with high-mass stars. We found no false positives in identifying LMXBs compared to the 32 positives in identifying HMXBs **(Figure 6)**. These misidentifications are likely due to the large number of low-mass systems in this dataset compared to high-mass systems. It was also possible to predict a companion star mass within 2.94 solar masses, with the temperature being the most significant contributor to the feature importance of the classifier **(Figure 4)**.

DISCUSSION

Current classification systems of XRBs would benefit from more specific classifications of the type of an XRB, as seen from the difference in accuracy in predicting mass and identifying XRB type from mass-related and non-mass-related factors. Since there is such a large discrepancy in accuracy in predicting these characteristics, it indicates that mass does not characterize these systems in a meaningful way (11). The classification of whether a system was high-mass or low-mass varied based on the specific database system used (12). Knowing the mass of a system does not allow one to precisely know much else about a system as opposed to a classification system based on something similar to a Hertzsprung-Russel diagram (13). Working off our results, we can now quantitatively classify XRBs into more specific categories.

We found a few limitations while analyzing this dataset. We started with a dataset of 517 XRBs, but that number was reduced to 110 due to missing variables. However, due to the already rare nature of these systems and the unavailability of data, we had to remove many from this dataset (14). Another concern we faced was conflicting data on certain variables;

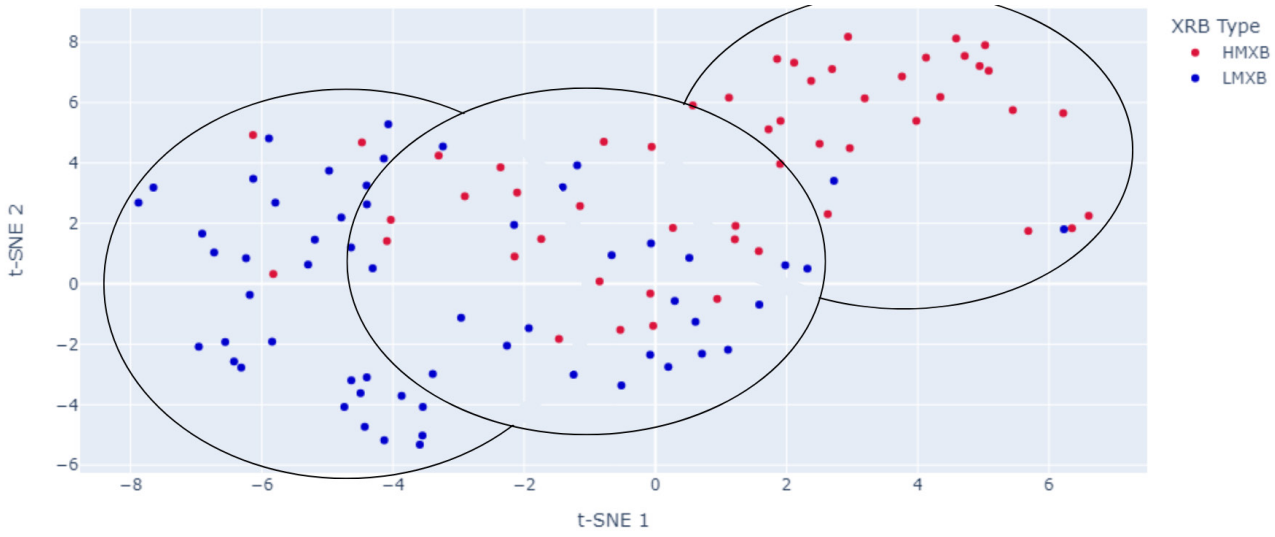

Figure 2. Divided t-SNE projection of XRB type. t-SNE plot showing high-mass (red) and high-mass (blue) X-Ray binaries, there is a divide in colors except for mild overlap in the middle. Utilized t-SNE entering mean distance, metallicity, orbital period, and effective temperature as a component, returning plot.

https://doi.org/10.59720/24-090

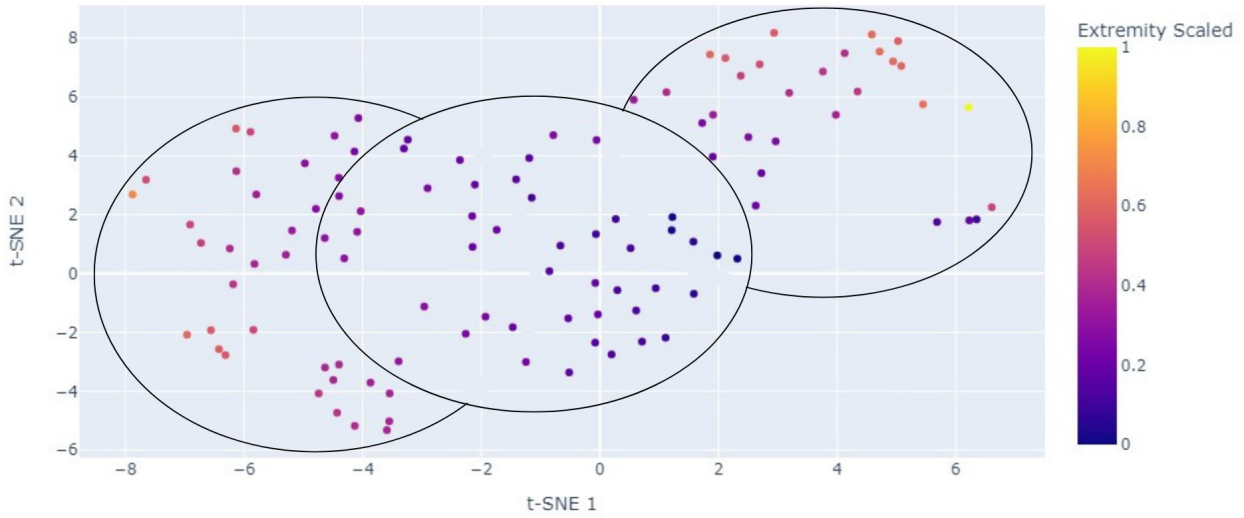

Figure 3. Colored t-SNE projection of XRB type. t-SNE of X-Ray binaries is colored by standardized effective temperature variance, with warmer colors correlating to a higher deviation away from mean temperature. There are three main temperature clusters correlating with overlapping orange and purple/pink dots, mainly blue dots, and overlapping orange and purple/pink dots again. We used t-SNE, entering mean distance, metallicity, orbital period, and effective temperature as a component of the t-SNE model, returning a plot with overlayed effective temperature.

cross-referencing multiple different stellar surveying archives and our original data, we found discrepancies in data in which we either had to throw out that data point or make an educated estimate. As machine learning is heavily dependent on the number of data points used, in which it is almost always better to have more data, we had to utilize specific techniques and methods that were able to deal with the smaller amount of data while sacrificing some amount of accuracy in speed in its analysis (9). We could utilize alternative methods, such as a neural network if we had more data for a more accurate analysis. This would be more accurate as neural networks allow for easier creation of non-linear models and can improve over multiple model generations (8,9).

A classification system like the Hertzsprung-Russell diagram, focusing on the temperature and metallicity of these systems, is a promising way to quantitatively classify XRBs while efficiently conveying much information about their properties. As mass increases, so does metallicity and higher temperatures. LMXBs show lower temperatures and metallicity. This aligns with accepted stellar evolution theories as more massive stars are typically more recently formed in the galactic plane, with higher metal density leading to a higher star's metallicity. Being able to reference temperature, metallicity, and mass would create a diagram that can effectively classify XRBs across multiple variables like the luminosity and temperature comparison in a Hertzsprung-Russel diagram in

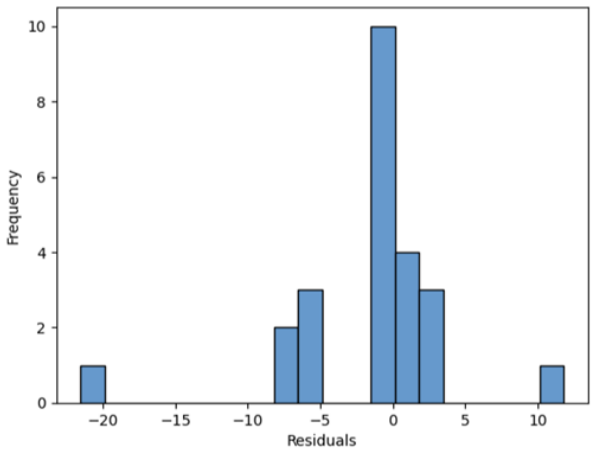

Figure 4. Comparison of effective temperature and metallicity of X-Ray binaries. Independent section of high-mass X-Ray binaries (red) generally only have metallicities greater than -1 with low-mass X-Ray binaries (blue) metallicities ranging from -4 to 1; there is overlap, as seen the same is in the t-SNE plot where these two groups intersect in temperature. We used Python libraries to plot data from a Pandas data frame.

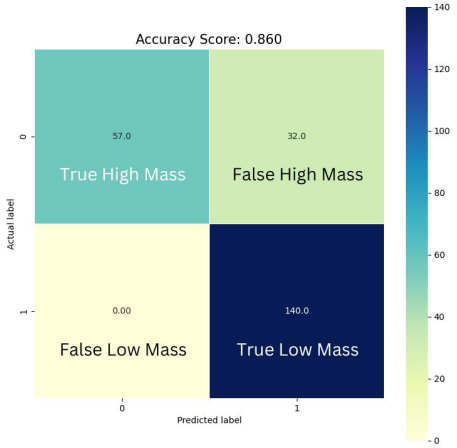

Figure 5. Residual histogram of mass prediction of X-Ray binaries. Error of predicted mass of X-Ray binaries typically near zero; however, in a few cases of severely off estimate, average mass is 9.06 solar masses with a range of 48.49 solar masses, with average mass error of 3.36. We used a logarithmic regression model taking into account distance, metallicity, orbital period, and effective temperature.

https://doi.org/10.59720/24-090

which star types are distinctly clustered. The next step is to fill in gaps in data to get a more precise classification of the XRBs.

We have found patterns that suggest a way to describe XRBs quantitatively. There are three distinct groups of temperatures, which can be utilized in the classification of these systems. As orbital velocity and pulsation period became more extreme, so did the temperature and metallicity. There is a high dependence on temperature and metallicity as defining characteristics of these systems. We have determined that it is possible to classify XRBs from their former broad classification of either HMXBs or LMXBs utilizing many different identifying characteristics, with the effective temperature of the system and the metallicity as the most significant markers of identification of these XRBs. We believe greater focus should be placed on creating a classification system that is based on these principles.

MATERIALS AND METHODS

Dataset

We obtained our primary dataset from the XRBCats High Mass XRB Catalog and the XRBCats Low Mass XRB Catalog. This dataset consisted of 517 XRBs, with 169 HMXBs and 348 LMXBs, each with data on right ascension, declination, mean distance (parsecs), pulsation period, orbital period, effective temperature, and mass (7, 15). From there, we added metallicity and redshift as factors pulling data from the Gaia Archive. After the dataset was reduced based on what data was missing, we had 110 XRBs. We decided on the XRBCats Catalog due to its incredibly up-to-date information and the vast amount of different data types within its dataset. We split the dataset into a training set, 70% of the data (77 data points), and a test set of 30% (33 data points).

Dimensionality reduction

To confirm that there is a pattern in the data that could be used to classify this dataset, we utilized t-SNE classification. We passed our dataset to a t-SNE algorithm, reducing the number of dimensions to 2. From there, we looked at the feature importance, seeing temperature and metallicity as large contributors. Several hyperparameters needed to be tuned. We chose a perplexity value of 30 to balance looking for local trends and global trends in the data (8). We also decided to include a random state value of 42 to reproduce the same results each time; by not using a random state value, the visualization may change each time it runs, the choice for 42 is arbitrary. We then overlaid standardized effective temperature data onto the t-SNE plot, seeing a correlation between extreme temperature values and extreme t-SNE values as displayed in transient black holes, an XRB (16).

Clustering algorithms

To cluster the data points, we used k-means clustering. We used Scikit-learn, an open-source machine-learning library in Python. We decided to use this library because there are many ML algorithms available (17). K-means clustering is an unsupervised centroid-based ML algorithm (9). This algorithm begins by randomly assigning data points a centroid, which will be the basis for defining the rest of the cluster (9).

ACKNOWLEDGMENTS

The authors want to thank the High School Research Academy program at The University of Texas at Austin for giving us the opportunity to conduct this research.

Received: April 5, 2024
Accepted: August 26, 2024
Published: October 9, 2025

REFERENCES

1. "X-Ray Binaries Monitoring." *Cosmos*, European Space Agency, www.cosmos.esa.int/web/cesar/x-ray-binaries-monitoring#:~:text=An%20X%2Dray%20binary%20consists,star%2C%20accretion%20rate%2C%20etc. Accessed 21 Nov. 2023.
2. "Black Holes and X-Ray Binaries." *Introduction to X-Ray Astronomy*, The Cambridge X-Ray Group, https://www-xray.ast.cam.ac.uk/xray_introduction/Blackholebinary.html. Accessed 22 Nov. 2023.
3. Fortin, Francis, et al. "A Catalogue of High-Mass X-Ray Binaries in the Galaxy: From the Integral to the *Gaia* Era." *Astronomy & Astrophysics*, vol. 671, 2023, https://doi.org/10.1051/0004-6361/202245236.
4. "X-Ray Binaries Reveal Extremes of Physics." *X-Ray Binaries Reveal Extremes of Physics*, California State Polytechnic University, Pomona, www.cpp.edu/sci/newsletter/x-ray-binaries-reveal-extremes-of-physics.shtml.
5. "Overview of Machine Learning." *Data Science Discovery*, University of Illinois Urbana-Champaign, https://discovery.cs.illinois.edu/learn/Towards-Machine-Learning/.
6. De Beurs, Zoe L., et al. "A Comparative Study of Machine-Learning Methods for X-Ray Binary Classification." *The Astrophysical Journal*, vol. 933, no. 1, 2022, p. 116, https://doi.org/10.3847/1538-4357/ac6184.
7. Avakyan, A., et al. "XRBcats: Galactic Low-Mass X-Ray Binary Catalogue." *Astronomy & Astrophysics*, vol. 675, July 2023, https://doi.org/10.1051/0004-6361/202346522.
8. "Visualization of Extremely High-Dimensional Neuroimaging and Phenotypic Data Using TensorBoard." *SOCR Visualization of Extremely High-Dimensional Neuroimaging Data*, https://socr.umich.edu/HTML5/SOCR_TensorBoard_UKBB/.
9. "Data Science and Predictive Analytics (UMich HS650)." *SOCR*, https://socr.umich.edu/people/dinov/2017/Spring/DSPA_HS650/notes/12_kMeans_Clustering.R.
10. Wilson-Hodge, Colleen A., et al. "*Nicer* and *Fermi* GBM Observations of the First Galactic Ultraluminous X-Ray Pulsar SWIFT J0243.6+6124." *The Astrophysical Journal*, vol. 863, no. 1, 2018, p. 9, https://doi.org/10.3847/1538-4357/aace60.
11. Liu, Q. Z., et al. "A Catalogue of Low-Mass X-Ray Binaries in the Galaxy, LMC, and SMC (Fourth Edition)." *Astronomy & Astrophysics*, vol. 469, no. 2, 2007, pp. 807–810, https://doi.org/10.1051/0004-6361:20077303.
12. Kaur, Ramanpreet, et al. "Multiwavelength Study of the Transient X-Ray Binary IGR J01583+6713." *Monthly Notices of the Royal Astronomical Society*, vol. 386, no. 4, 2008, pp. 2253–2261, https://doi.org/10.1111/j.1365-2966.2008.13233.x.
13. "Field Guide to X-Ray Astronomy :: Binary and Multiple Star Systems." *Chandra*, https://chandra.harvard.edu/

https://doi.org/10.59720/24-090

xray_sources/binary_stars.html.
14. Grunhut, J. H., et al. "Orbit and Properties of the Massive X-Ray Binary BD+6073=IGR J00370+6122." *Astronomy & Astrophysics*, vol. 563, 2014, https://doi.org/10.1051/0004-6361/201322738.
15. Neumann, M., et al. "XRBCATS: Galactic High Mass X-Ray Binary Catalogue." *Astronomy & Astrophysics*, vol. 677, 2023, https://doi.org/10.1051/0004-6361/202245728.
16. Kara, E., et al. "The Corona Contracts in a Black-Hole Transient." *Nature*, vol. 565, no. 7738, 2019, pp. 198–201, https://doi.org/10.1038/s41586-018-0803-x.
17. "Learn." *Scikit*, https://scikit-learn.org/stable/.