Machine Learning the Spectrum of X-ray Binaries

Nom de l'alumne: Elizabeth Birch Hardwick

Grau: Física

Nom de l'empresa o Institució on ha realitzat l'estada: Institute of Cosmology and Exoplanetary Science (ICE CSIC)

Nom del Tutor: Dr. Konstantinos Kovlakas

Període de temps en que s'ha fet l'estada: 10/2024 - 5/2025

Data d'entrega de l'informe: 23/06/2025

1 Introduction to the Host Institution

The Institute of Space Sciences (ICE, CSIC) is a research center located near Barcelona and part of the Spanish National Research Council. It conducts cutting-edge research in astrophysics, cosmology, planetary systems, and space instrumentation. ICE is involved in major international collaborations such as Gaia, Euclid, and LISA, and also contributes to the development of scientific instrumentation through the Institute of Space Studies of Catalonia (IEEC). Among its research lines is the study of compact objects and high-energy astrophysics, including X-ray binaries, which formed the focus of my internship project.

2 Introduction

X-ray binaries (XRBs) are binary star systems consisting of a compact object, typically a neutron star or a black hole, and a companion star. These systems generate intense X-ray emissions due to the accretion of material from the companion onto the compact object. This accretion process forms a hot disk of gas, which radiates predominantly in the X-ray spectrum. XRBs are classified into high-mass XRBs (HMXBs), associated with young, massive stars, and low-mass XRBs (LMXBs), linked to older stellar populations. Their prevalence and brightness in galaxies make them essential for studying the X-ray output of normal galaxies and understanding high-energy astrophysical processes.

The X-ray spectrum of these binaries is particularly intriguing because it encodes critical information about the physical processes at work, such as accretion physics and interactions between matter and radiation. The intrinsic spectra of XRBs are often described using a power-law model, given by:

$$\Phi(E) = KE^{-\Gamma},$$

where $\Phi(E)$ is the photon flux (in units of $\mathrm{cm^{-2}\ s^{-1}\ keV^{-1}}$), K is the normalization constant, E is the photon energy in keV, and Γ is the photon index, which determines the slope of the spectrum. Another important parameter is the hydrogen column density, N_{H}, which quantifies the absorption of soft X-rays along the line of sight, with units of cm^{-2}. This means that for the observed spectrum

$$\Phi_{obs}(E, N_{\mathrm{H}}) = A(E, NH)\Phi(E)$$

where A is a function that determines the fraction of photons of energy E absorbed for a given hydrogen column density, an accurate determination of N_{H} and Γ is essential for reconstructing the intrinsic spectral shape of XRBs, through a fitting process.

However, deriving N_{H} and Γ possesses significant challenges. Factors such as the intrinsic variability of XRBs, limited photon counts in observations, and instrumental biases complicate the fitting process. Additionally, the distances to

XRB-hosting galaxies introduce further uncertainties. A promising alternative to direct spectral fitting is the use of X-ray colours, which provide a simplified representation of the spectral shape and offer a practical solution for characterising XRB populations, particularly in low photon count settings (e.g., faint and/or distant sources).

X-ray colours are defined as ratios of photon counts in specific energy bands. For this study, the soft colour (C_{S}) and hard colour (C_{H}) are expressed as:

$$\mathrm{soft} = \frac{M-S}{M+S}, \mathrm{hard} = \frac{H-M}{H+M},$$

where S, M, and H are the photon counts in the soft (0.5–1 keV), medium (1–2 keV), and hard (2–7 keV) bands, respectively. These colour ratios serve as proxies for spectral parameters and are less affected by observational uncertainties compared to direct spectral fitting.

2.1 Objectives

In this study we have three main objectives:

1. Investigate whether the X-ray colours measured by the *Chandra X-ray Observatoy* can be used to infer the spectral parameters of X-ray sources.

2. Find which machine-learning method is most accurate for the task.

3. Study whether the degradation of the optics in *Chandra* affect the inference of the parameters.

3 Methodology

3.1 Data and visualisation

To study the appropriate machine learning framework for predicting N_{H} and Γ values (Objectives 1 and 2), we need a dataset with a wide range of values for these parameters. We selected N_{H} values uniformly distributed in log space between $10^{19.5}$ and $10^{23.5}$, and Γ values uniformly distributed between 0 and 5. Through collaborators (Izabela Pavel; private communication), we received simulated spectra corresponding to the sampled parameters for two observational cycles (Objective 3). These cycles correspond to the initial (ObsCycle 3) and current (ObsCycle 25) stages of the *Chandra X-ray Observatory*, allowing us to probe the effect of instrumental degradation.

To visualise the data we created Figure 1 As we can see for most of the values, it follows a grid-like structure as in Figure 2 except for the edge of high soft band values, where the gamma values are high regardless of the gird and the "tail" at high hard band values where NH values are high. We can also appreciate that there are differences between the two cycles, especially in the soft band where Cycle 25 doesn't have negative values as seen in Figure 3. This is expected, as the detector's efficiency decreases over time due to molecular contamination

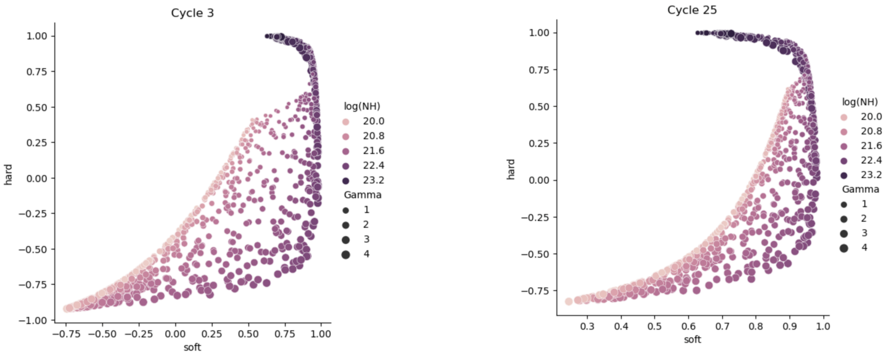

Figure 1: Distribution of the soft and hard colours of our Cycle 3 (left) and Cycle 25 (right) data, with the corresponding $\log(N_H)$ and Γ values represented by the points' colour and size, respectively.

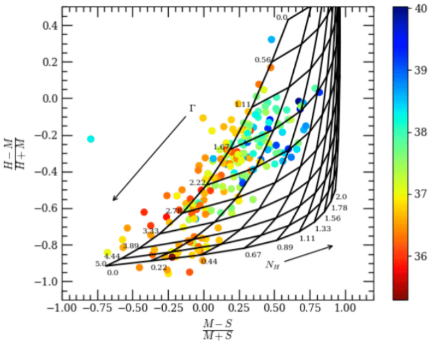

Figure 2: Distribution of simulated soft and hard colours with the grid showing the relationship between NH and Γ and the colour representing their decimal logarithmic of luminosity. Adapted from Pavel et al. [1]

on optical blocking filters and changes in charge transfer efficiency, reducing sensitivity to soft photons and shifting measured flux values.

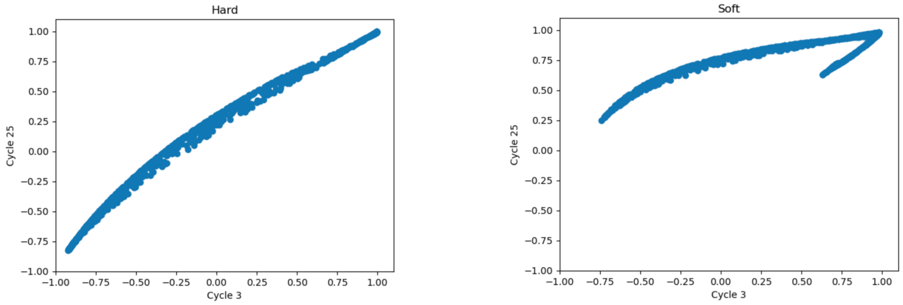

Figure 3: Distribution of the hard (left) and soft (right) colours in the cycle 25 vs cycle3 space.

3.2 Machine learning approach

To predict the NH and Γ values from the hard and soft colour bands, we evaluated the performance of three different machine learning models: Polynomial Regression, KNN Regression, and Random Forest Regression.

To assess the effectiveness of these models, we divided the data into two sets:

- **Training set:** Used to train the model, enabling it to learn patterns from the data.
- **Testing set:** Used to evaluate the model's performance on unseen data, providing an unbiased estimate of its generalization ability.

However, a single train-test split can introduce variability in the results, as the specific division of the data might influence model performance. For instance, some splits may inadvertently place more difficult samples in the testing set, leading to inconsistent or unreliable evaluations. To ensure this doesn't happen we implemented 5-fold cross-validation:

1. The training dataset is divided into 5 equal parts (folds).
2. The model is trained on 4 folds and tested on the remaining fold.
3. This process is repeated 5 times, with each fold serving as the testing set exactly once.
4. The performance metric is averaged across the 5 iterations, providing a more stable and reliable evaluation for hyperparameter selection.
5. The optimized model is then applied on the testing set giving us the final performance metric.

Since each model has it's own hyperparameters and we don't know which ones yield the best results we implemented hyperparameter tuning using the R^2 score to find the best ones thus reducing the chance of overfitting or underfitting.

3.2.1 Polynomial regression

Since the data has no significant outliers (apart from perhaps the "tail") and appears to follow a polynomial-like structure, polynomial regression seems to be the most natural starting point. Polynomial regression is an extension of linear regression where the relationship between the input variables and the target variable is modelled as an n-degree polynomial. Instead of fitting a straight line to the data, this approach allows for the flexibility of curves, which can better capture non-linear relationships in the dataset.

The key hyperparameter to tune in this method is the degree of the polynomial. The degree determines the complexity of the curve: lower degrees (e.g., 1 or 2) produce simpler models that may underfit the data, while higher degrees allow the model to adapt to more intricate patterns but risk overfitting by capturing noise rather than meaningful trends. Finding the right balance is essential to ensure the model generalizes well to unseen data.

3.2.2 KNN regression

The problem with polynomial regression is that it's not very good at extrapolation. When the model is trained on a specific range of data, it struggles to make accurate predictions for data points outside this range, as the polynomial curve can sharply deviate beyond the training data. To make sure that this wasn't causing worse performance, we tested a KNN regression method.

K-Nearest Neighbors (KNN) regression is a non-parametric method that makes predictions based on the average of the k-nearest data points to a given test point. The idea is that similar data points should have similar output values, so the prediction is made by looking at the closest points in the feature space. In this method, the key hyperparameter to tune is the number of nearest neighbors (k), which controls how many surrounding points contribute to the prediction. A smaller value of k can lead to a model that is sensitive to noise (overfitting), while a larger value smooths the prediction but may underfit the data by not capturing enough local structure.

3.2.3 Random Forrest regression

With the KNN method, there is one drawback: it's based on a definition of the distance between points in the feature space. However, when the features are not of the same units (e.g., lengths), or they are transformed (e.g., logarithmic quantities), an appropriate distance measure needs to be explored, weighting the contribution of the different features. For this reason we also explore the Random Forest technique that does not depend on any distance measure, but the ordinal relationship of feature values. It works by constructing multiple decision trees during training, each trained on a random subset of the data, and then averaging their predictions to make a final decision, as can be seen in Figure 4. This approach allows Random Forest to capture complex, non-linear relationships in the data without being affected by the scale of the features.

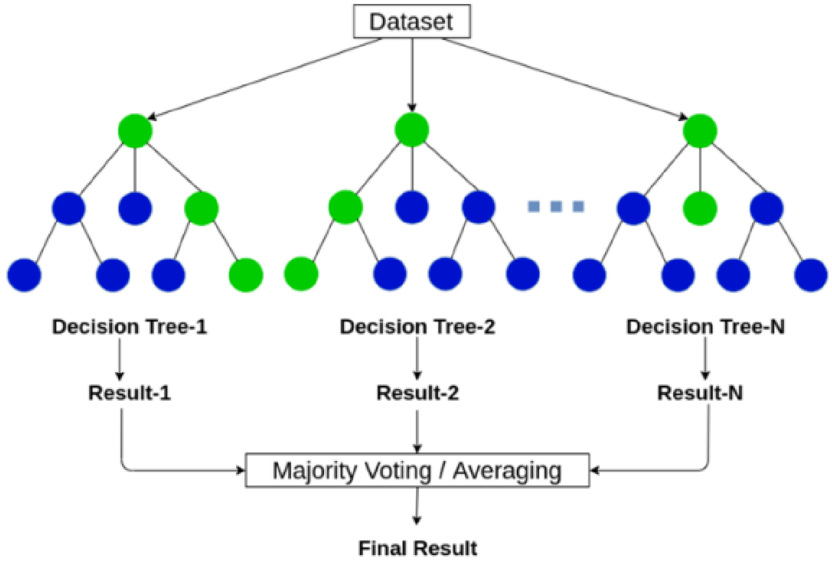

Figure 4: Diagram that shows how Random Forrest regression works.

In this method,the two key hyperparameters to tune on are:

- **n_estimators:** This is the number of trees in the forest. A good starting point is 100 trees, but experimenting with values like 300 or 1000 can help determine the effect of increasing the number of trees on the model's performance.
- **max_depth:** This controls how deep each tree can grow. Limiting the depth of the trees helps prevent overfitting. We explored values such as 5, 10, 20, and None (which allows the trees to grow until each leaf contains a single data point or no further splits improve the prediction).

By fine-tuning these parameters, Random Forest can adapt to the data's underlying structure and learn both broad patterns and finer, more localized details, providing a more robust and accurate prediction model.

4 Results

The full code used for these analyses can be found in my GitHub repository [6].

4.1 Comparison of the models

To compare the three models, we utilized the data from Cycle 3, as it contained both positive and negative values for the soft band, offering a more balanced dataset for evaluation.

During hyperparameter tuning for each model, we identified the optimal parameters that yielded the best performance. For Polynomial Regression, the best results were obtained with polynomial degrees in the range of 9–11. For KNN Regression, the optimal number of neighbours was found to be 2–3. Similarly, for Random Forest Regression, the most effective hyperparameters were a maximum tree depth of approximately 20 and around 800 estimators.

For the subsequent graphs and results, we used the following hyperparameter values: degree = 11, n_neighbours = 2, depth = 20, and n_estimators = 800.

In Figure 5, the predicted values of $\log(N_H)$ demonstrate a high degree of accuracy, with the Polynomial Regression model showing results that are nearly exact. For KNN and Random Forest Regression, a slight deviation is observed at lower values of $\log(N_H)$.

For Γ, all three models produce predictions that are generally accurate, although a few points exhibit noticeable variation.

From the contour maps of *Mean Absolute Error (MAE)* as seen in Figure 6, Polynomial Regression performs well across most of the parameter space but shows higher errors near the boundaries, particularly at large values of $\log(NH)$ and extreme values of Γ. KNN Regression achieves generally low error but is more sensitive to local variations, with noticeable high-error regions around $\Gamma \approx 0.5$–1.0 and $\log(NH) \approx 22.0$. Random Forest Regression demonstrates the most consistent performance, maintaining low errors across the entire space with fewer and less severe high-error regions.

Finally, in Table 1 are some examples of the R^2 scores for different train-test splits.

Train-Test Split	**Polynomial**	**KNN**	**RFR**
1	0.9190	0.9061	0.9303
2	0.9223	0.8812	0.9119
3	0.9435	0.9102	0.9337
4	0.9298	0.9236	0.9448
5	0.9385	0.8759	0.8988

Table 1: Performance scores for different train-test splits and regressors.

The Random Forest Regressor consistently outperforms the KNN Regressor. While the Polynomial Regressor shows greater consistency in performance, the RFR occasionally achieves better results, highlighting its potential for specific cases.

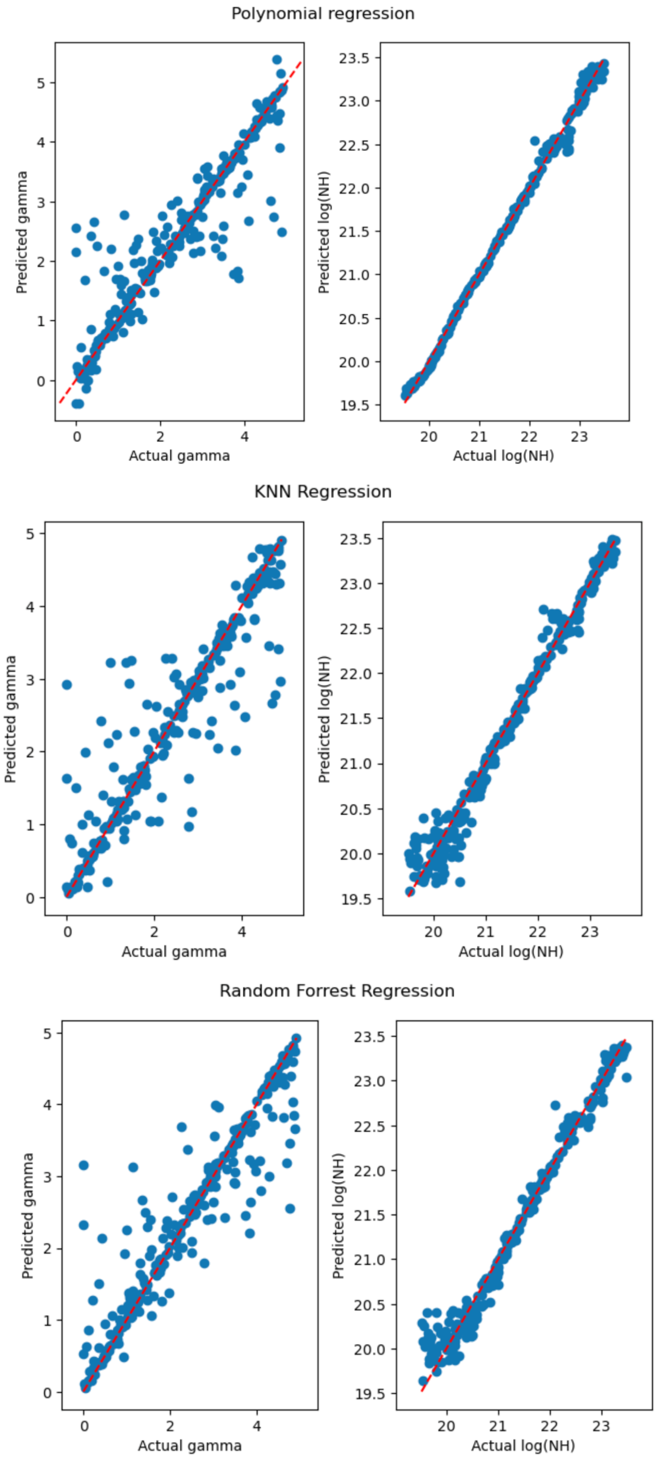

Figure 5: Predicted vs. actual values for $\log(N_H)$ (left) and Γ (right) for the three regression models tested.

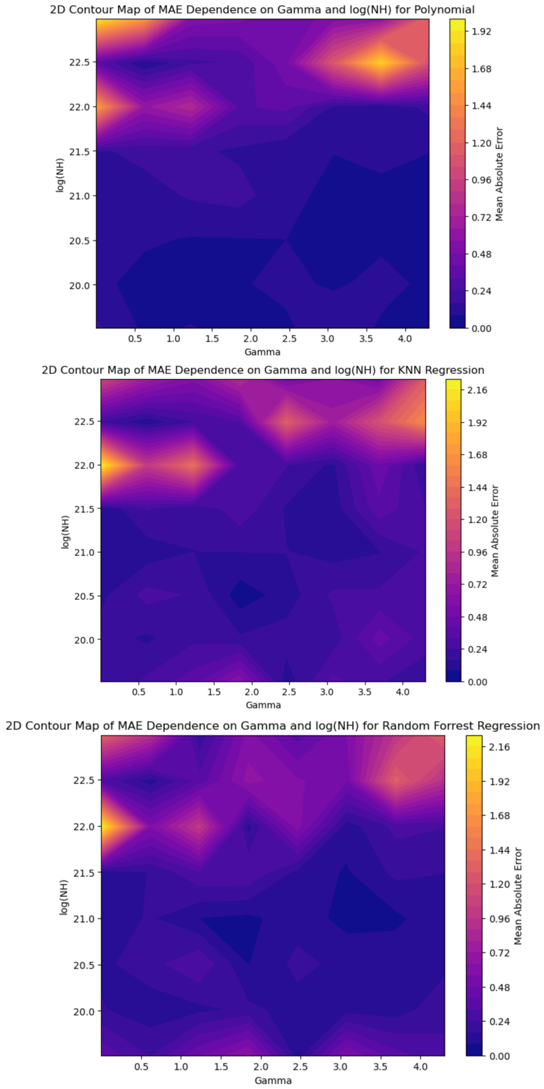

Figure 6: Contour maps of the Mean Absolute Error (MAE) for Polynomial Regression (top), KNN Regression (middle), and Random Forest Regression (bottom).

4.2 Testing on different ObsCycles

To test whether the Obscycle affects the relationships, we trained two models on Cycle 3 and Cycle 25, respectively, and tested them on the other cycle's data using the polynomial regression method.

Model and Dataset	R^2	**MAE**
Model 3 on Dataset 3	0.9339	0.1545
Model 25 on Dataset 25	0.9340	0.1885
Model 3 on Dataset 25	0.5013	0.5546
Model 25 on Dataset 3	-1922304968717.86	583262.8516

Table 2: Performance of polynomial models on different datasets using R^2 and MAE metrics.

The results, as illustrated in Table 2, indicate that the models are not accurate when applied to different cycles. The magnitudes of this outcome is unexpected, as one would anticipate that using Model A on Dataset B, and vice versa, would produce errors of the same order of magnitude in terms of MAE or R^2.

To investigate the significant variations in magnitudes, we generated Figure 7. For Model 3 applied to Cycle 25, the gamma predictions are reasonably accurate overall; however, for lower values of $\log(NH)$, the model consistently overestimates the predictions. Conversely, when Model 25 is applied to Cycle 3, the predictions deviate substantially, with several values falling outside the expected range. Although the model appears to perform adequately for low gamma and high $\log(NH)$ values, a closer examination of a zoomed-in version of the graph reveals a similar pattern of discrepancies. In this case, the model tends to underestimate values, mirroring the behaviour observed with Model 3 on Cycle 25.

To determine whether the observed differences in magnitudes were due to extrapolation, we conducted an analysis using the KNN regressor:

Model and Dataset	R^2	**MSA**
Model 3 on Dataset 3	0.9205	0.1924
Model 25 on Dataset 25	0.8999	0.2354
Model 3 on Dataset 25	0.4763	0.5786
Model 25 on Dataset 3	0.3343	0.8053

Table 3: Performance of KNN models on different datasets using R^2 and MAE metrics.

As shown in Table 3, the KNN regressor continues to perform poorly across different observation cycles, indicating the necessity of using a distinct model for each ObsCycle.

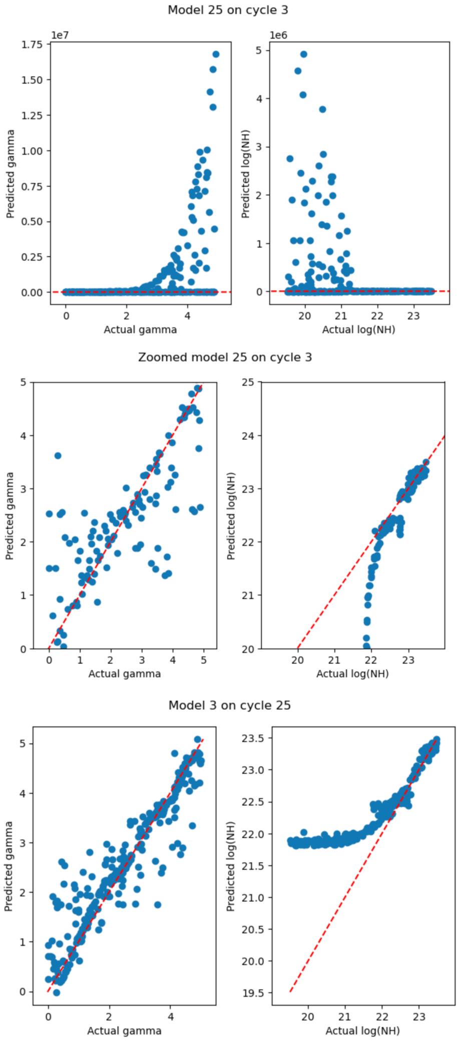

Figure 7: Predicted vs actual values of Γ (left) and $\log(N_H)$ (right) when applying models trained on one ObsCycle to data from another cycle.

4.3 Testing without the "tail"

Upon analysing the points with high MAE, as shown in Figure 8, it was observed that the majority of problematic points were located in the tail or at the edge with high soft colour values.

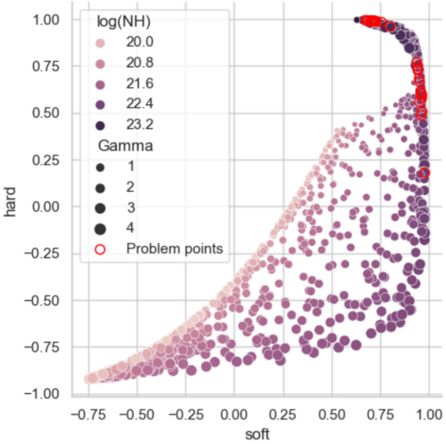

Figure 8: Locations of the data points with the highest Mean Absolute Error (MAE) in the colour-colour space.

To address this, a polynomial regression model was trained on a dataset excluding the tail (defined as hard > 0.6) and edge (defined as soft > 0.9) regions, as shown in Figure 9.

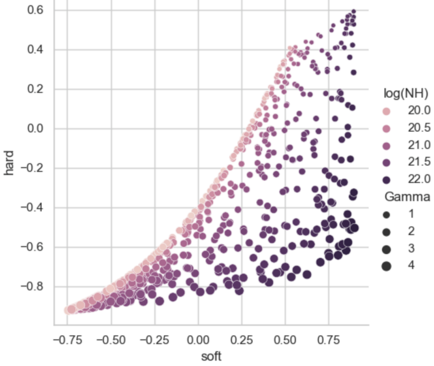

Figure 9: Soft and hard colour distribution after removing the "tail" (hard $>$ 0.6) and edge (soft $>$ 0.9) regions from the dataset.

The results, illustrated in Figure 10, indicate that the modified model performed significantly better yielding a near-perfect performance with $R^2 = 0.9984$ and $MAE = 0.0141$.

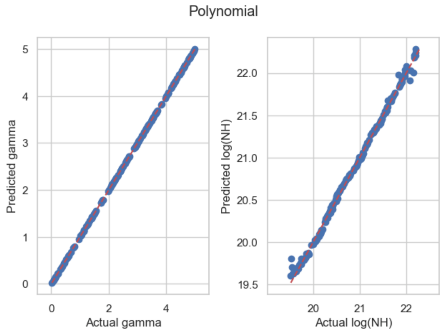

Figure 10: Predicted vs actual values of $\log(N_H)$ (left) and Γ (right) for the Polynomial Regression model after removing the “tail” and edge regions.

The improved performance observed when excluding the tail and edge regions can be attributed to the relationship between increasing Γ or N_H and the resulting spectral properties. Specifically, as Γ or N_H increase, the spectrum becomes either excessively ”soft” or ”hard,” leading to a loss of photons at the lower and higher ends of the energy range. In scenarios with low photon counts, this results in large uncertainties in the measured counts and the derived colours. It is important to note that these are not ideal spectra (e.g., those measured directly next to the source) but are instead outputs from simulations that predict telescope observations. Additionally, the reduced sensitivity of the detector at lower energies leads to the loss of many soft photons. Consequently, numerous spectra begin to resemble each other, and the colour measurements lose their ability to effectively distinguish between them.

5 Conclusions

The results confirm that machine learning provides a reliable approach for predicting the spectral parameters N_H and Γ from X-ray colour data. The analysis shows a clear relation between the colour bands and these parameters, allowing for accurate predictions without requiring direct spectral fitting.

Among the models tested, Random Forest Regression (RFR) consistently achieved better performance than KNN Regression and performed similarly or better than Polynomial Regression. The ability of RFR to model non-linear relations while remaining stable across different regions of the parameter space makes it a reliable choice. However, the Polynomial Regression model also performed well and remains a valid alternative, especially since it’s faster and allows us to extract the parameters from the trained model. These parameters can then be easily applied to other values, as demonstrated in the GitHub repository [6], where the process of loading the model and utilizing its parameters for further predictions is clearly outlined.

One of the main findings is that the model’s performance depends on the observational cycle. Models trained on one cycle do not generalize well to another, highlighting the necessity of training separate models for different cycles. This result is consistent with the expected changes in detector response over time, reinforcing the importance of accounting for instrumental degradation when applying machine learning to X-ray data.

Additionally, removing the “tail” and extreme soft band values significantly improved performance, indicating that these regions introduce larger uncertainties that affect performance. This suggests that future implementations should either train separate models for these problematic regions.

Overall, this study shows that machine learning can successfully estimate spectral parameters from X-ray colours, providing a practical alternative in cases where traditional spectral fitting is limited by low photon counts or instrumental effects.

6 Personal Conclusions

This internship at the Institute of Space Sciences (ICE, CSIC) has been a highly valuable experience. Working on the classification of X-ray binary spectra using machine learning has allowed me to apply my knowledge of physics in a real-world, computational setting. It has also given me the opportunity to combine my passion for astrophysics with the programming and data analysis skills I’ve developed through my studies and additional coursework in machine learning and data science. This experience has deepened my interest in computational methods for solving physical problems and has reinforced my desire to continue in this field in my future academic and professional career.

I am especially grateful to my tutor, Dr. Konstantinos Kovlakas, for his continuous guidance, support, and availability throughout the project. His feedback and expertise have been invaluable to my development. Overall, this internship has been both academically enriching and personally fulfilling, and has provided me with a clearer vision of the type of research I would like to pursue in the future.

References

[1] I. Pavel, K. Kovlakas, B. D. Lehmer, *The spectral shape of X-ray binary populations in nearby galaxies*, Master Thesis, Autonomous University of Barcelona, 2024.

[2] F. Seward and P. Charles, *Exploring the X-ray Universe*, 2nd ed., Cambridge University Press, 2010.

[3] J. Andrews, P. Bonfini, K. Kovlakas, G. Maravelias, “2024 Summer School for Astrostatistics in Crete”, *GitHub Repository*, 2024, `https://github.com/astrostatistics-in-crete/2024_summer_school`.

[4] Scikit-learn developers, “Scikit-learn Documentation”, *Scikit-learn*, 2024, `https://scikit-learn.org/stable/`.

[5] NASA, “Chandra X-ray Observatory”, *NASA*, 2024, `https://cxc.harvard.edu`.

[6] E. Birch Hardwick, “Machine Learning the Spectrum of X-ray Binaries”, *GitHub Repository*, 2025, `https://github.com/elizabethbirchhardwick/Machine_Learning_the_Spectrum_of_X-ray-Binaries/tree/main`.