Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **524,** 4801–4818 (2023)
Advance Access publication 2023 June 09

https://doi.org/10.1093/mnras/stad1643

QPOML: a machine learning approach to detect and characterize quasi-periodic oscillations in X-ray binaries

Thaddaeus J. Kiker [1,2,3]★ James F. Steiner,[4] Cecilia Garraffo,[4] Mariano Méndez [5] and Liang Zhang [6]

[1] *Sunny Hills High School, 1801 Lancer Way, Fullerton, CA 92833, USA*
[2] *Department of Physics, Columbia University, New York, NY 10027, USA*
[3] *Earth Science Division, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA*
[4] *Center for Astrophysics | Harvard & Smithsonian, 60 Garden St. Cambridge, MA 02138, USA*
[5] *Kapteyn Astronomical Institute, University of Groningen, P.O. BOX 800, NL-9700 AV Groningen, the Netherlands*
[6] *Key Laboratory for Particle Astrophysics, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China*

Accepted 2023 April 24. Received 2023 April 24; in original form 2022 November 21

ABSTRACT

Astronomy is presently experiencing profound growth in the deployment of machine learning to explore large data sets. However, transient quasi-periodic oscillations (QPOs) that appear in power density spectra of many X-ray binary (XRB) system observations are an intriguing phenomena heretofore not explored with machine learning. In light of this, we propose and experiment with novel methodologies for predicting the presence and properties of QPOs to make the first ever detections and characterizations of QPOs with machine learning models. We base our findings on raw energy spectra and processed features derived from energy spectra using an abundance of data from the *NICER* and *Rossi X-ray Timing Explorer* space telescope archives for two black hole low-mass XRB sources, GRS 1915+105 and MAXI J1535−571. We advance these non-traditional methods as a foundation for using machine learning to discover global inter-object generalizations between – and provide unique insights about – energy and timing phenomena to assist with the ongoing challenge of unambiguously understanding the nature and origin of QPOs. Additionally, we have developed a publicly available PYTHON machine learning library, QPOML, to enable further machine learning aided investigations into QPOs.

Key words: accretion, accretion discs – black hole physics – stars: individual (GRS 1915+105, MAXI J1535+571) – X-rays: binaries.

1 INTRODUCTION

At the ends of their lives, massive stars 'do not go gentle into that good night' (Thomas 1952). Instead, if their initial mass exceeds $\sim 8\ \mathrm{M}_{\odot}$, core-collapse leads to spectacular Type II supernovae (Schlegel 1995). If the compact remnant remains bound or becomes bound to a non-degenerate companion star, the result can be a neutron star (NS) or black hole (BH) remnant (Gilmore 2004). In special cases, this object maintains a non-degenerate partner, and together these may form an X-ray binary (XRB) system, in which the non-degenerate star engages in mass-exchange with its compact partner (Tauris & van den Heuvel 2006). Such systems are characterized by accretion from the donor star, through accretion discs (Shakura & Sunyaev 1973) and are the sources for jets (Gallo, Fender & Kaiser 2005; van den Eijnden et al. 2018) and winds (Neilsen 2013; Castro Segura et al. 2022). Additional exotic phenomena like thermonuclear surface burning (Bildsten 1998) have also been observed in NS binaries. Both BH and NS systems are both observed to emit thermal X-ray radiation with temperatures ∼1 keV that is understood to arise from the conversion of gravitational potential to radiative energy. NSs can produce thermal emission at their surfaces, and the optically thick, geometrically thin accretion discs around both NSs and BHs can produce strong thermal X-ray emission (Shakura & Sunyaev 1973). Furthermore, BH and NS XRBs both also show hard X-ray flux coming from Compton up-scattering of thermal disc emission by a cloud of hot electrons around the compact source known as the corona (Galeev, Rosner & Vaiana 1979; White & Holt 1982). Comptonized emission is commonly modelled by a power-law relationship $N(E) \propto E^{-\Gamma}$, where Γ is the photon index (McClintock & Remillard 2006). Strongly Comptonized spectra commonly exhibit reflection features like a fluorescent, relativistically broadened 6.4 keV Fe Kα line (Fabian et al. 1989) and ∼30 keV Compton hump (Ross & Fabian 2005). These systems can be transient in activity and undergo evolution in spectral states (Gardenier & Uttley 2018), ranging from hard, to intermediate, and to soft (McClintock & Remillard 2006), which are coupled with mass-accretion rate (Done & Gierliński 2004), spectral hardness or thermal dominance, and thereby position on a hardness–intensity or colour–colour diagram track (Ingram & Motta 2019), and the presence/absence of quasi-periodic oscillations (QPO) of the observed X-ray radiation (McClintock & Remillard 2006). These QPOs are detected as narrow peaks in power-density spectra (PDS; Homan & Belloni 2005). In the past 30 yr, numerous theories, including but not limited to relativistic precession (Stella & Vietri 1998), precesssing inner flow (Ingram, Done & Fragile 2009), corrugation modes (Kato & Fukue 1980), accretion ejection instability (Tagger & Pellat 1999), and propagating oscillatory shock (Molteni, Sponholz & Chakrabarti 1996) have been advanced to

★ E-mail: thaddaeuskiker@protonmail.com

explain the occurrence of QPOs in BH, as well as NS, XRB systems. Yet, there is not consensus as to which model is most plausible. In black-hole systems, most of the observed QPOs have been at low frequencies (LF) ≤30 Hz (Belloni et al. 2020). Only a small subset has BHXRBs have exhibited high-frequency QPOs (HFQPO). LF QPOs are further subdivided canonically into three classes (Casella, Belloni & Stella 2005): Type-A QPOs are the rarest, sometimes appearing in the intermediate or soft state as broad, low amplitude features centred between 6 and 9 Hz and usually lacking harmonic companions (Motta et al. 2011). Type-B QPOs are more common, and can be seen during the short soft intermediate state and have shown some connection with jet behaviour (Gao et al. 2017; García et al. 2021). Finally, type C QPOs are the most common, and can be detected as narrow features in the low-hard and hard-intermediate states with harmonic companions (Fragile, Straub & Blaes 2016). Their fundamental frequencies range from ∼0.1 to 30 Hz depending on state, and almost always correlate strongly with spectral features like Γ and luminosity (Motta et al. 2015). As for HFQPOs, we recommend readers to Motta et al. (2011), Méndez et al. (2013), and Stella & Vietri (1999). QPOs are also observed in NS systems (Belloni, Psaltis & van der Klis 2002; Wang 2016). We focus on LFQPOS from BHXRBs in this paper and recommend van der Klis (2006) and Wang (2016) for reviews of NS specific QPOs and Ingram & Motta (2019), Jonker, van der Klis & Wijnands (1999), Kato (2005), Revnivtsev et al. (2001), and Méndez & Belloni (2021) of QPOs in XRBs in general. All in all, hundreds of XRBs have been observed since the discovery of Sco X-1 (Giacconi et al. 1962; Liu, van Paradijs & van den Heuvel 2007; Corral-Santana et al. 2016) and a large fraction show some type of QPO.

Machine learning is a revolutionary subfield of artificial intelligence in which models teach themselves patterns in data rather than operating by externally supplied hard-coded rules (Goodfellow, Bengio & Courville 2016). With data available to astronomers approaching the petabyte domain (Ivezić et al. 2014), this aspect of machine learning has helped it supplement traditional methods in addressing the ever growing volume and increasing complexity of astronomical data, while also providing new perspectives on old phenomena (Kremer et al. 2017; Rodríguez, Rodríguez-Rodríguez & Woo 2022). Consequently, machine learning has been used prolifically to classify variable stars (Richards et al. 2011), search for exoplanets (Pearson, Palafox & Griffith 2018), detect pulsars (Zhu et al. 2014), predict solar flares (Li et al. 2020), classify and even discover galaxies (Dieleman, Willett & Dambre 2015; Kojima et al. 2020). However, although machine learning techniques has been applied to a number of problems related XRBs as well, e.g. to classify and identify X-ray binaries (Huppenkothen et al. ; Arnason, Barmby & Vulic 2020; Sreehari & Nandi 2021; de Beurs et al. 2022; Orwat-Kapola et al. 2022; Yang et al. 2022b), predict compact object identity (Pattnaik et al. 2021), and study gravitational waves (Schmidt et al. 2021), this subfield contains tens of thousands of observations that have never been explored with machine learning to detect QPOs themselves. For the first time, in this work we seek to develop a methodology for using machine learning to detect QPOs, because we believe that our theoretical understanding of QPOs and their exotic progenitor systems would benefit from insights this approach could provide (Fudenberg & Liang 2020). Our approach is unique, because although the externally determined presence of QPOs has been used as a binary input parameter in accretion state classifiers such as those in Sreehari & Nandi (2021), QPOs have never before been the output of machine learning prediction themselves. The rest of this paper is structured as follows: in Section 2 we describe the observations upon which we base our work. Following this, in Section 3 we describe the energy and spectral fitting procedures we employ to produce input/output data from these observations for the machine learning models and methods which we detail in Section 4. We present our results in Section 5, and we discuss these results contextually in Section 6. Finally, we conclude in Section 7. Additional work concerning demonstrating QPOML and model comparison are presented in following appendices.

2 OBSERVATIONS

2.1 GRS 1915+105

GRS 1915+105 is a well-studied galactic low mass XRB system composed of a $12.4^{+2.0}_{-1.8}$ M$_\odot$ primary and a 1.2 M$_\odot$ K III secondary (Greiner et al. 2001; Greiner 2003) on a 34 d period located at a distance of $8.6^{+2.0}_{-1.6}$ kpc from the Earth (Reid et al. 2014). The secondary star in this system overflows its Roche lobe. GRS 1915+105 was one of the first microquasar jet systems, with (apparent) superluminal motion detected from a ballistic jet launched with an inclination 70 ± 2 deg (Mirabel & Rodríguez 1994). Since its discovery in 1992 (Castro-Tirado, Brandt & Lund 1992), this somewhat peculiar source has displayed unique timing and spectral patterns which have been organized into 14 separate variability classifications depending on its variability state (Belloni et al. 2000; Hannikainen et al. 2005). With its 16-yr archive of observations of this source we considered all data from the *Rossi X-ray Timing Explorer* (RXTE) Proportional Counter Array (PCA; $2-60$ keV) that are also included in Zhang et al. (2020), Méndez et al. (2022), and García et al. (2022a). These include a great number of detections of type C QPOs between 1996 and 2012. Energy and PDS have been derived from binned, event, and GoodXenon data as described in Zhang et al. (2020). Briefly, PDS have been constructed by averaging 128 s long intervals at 1/128 s time resolution, normalized according to Leahy, Elsner & Weisskopf (1983), and Poisson noise subtracted (Zhang et al. 1995). Of the 625 timing observations in Zhang et al. (2020), we have 554 matching energy spectra.

2.2 MAXI J1535−571

MAXI J1535−571 was discovered by the MAXI/GSC nova alert system as a hard X-ray transient system undergoing outburst in 2017 by Negoro et al. (2017a), and it was first suggested to be BH system by Negoro et al. (2017b). Since discovery, it has been suggested as an ∼10.39 M$_\odot$ BH, ∼5 kpc distant (Sridhar et al. 2019). MAXI J1535−571 has displayed state transitions (Nakahira et al. 2018), reflaring events (Cúneo et al. 2020), and hysteresis during its main outburst (Parikh et al. 2019). Furthermore, it has been determined to possess a near-maximal dimensionless spin parameter of $a = \frac{cJ}{GM^2} > 0.99$ (Miller et al. 2018; Liu et al. 2022). To study this source we use data from the International Space Station mounted, soft X-ray (0.5–12 keV) observatory Neutron star Interior Composition ExploreR (NICER; Gendreau, Arzoumanian & Okajima 2012) which has unequaled spectral-timing capabilities in soft X-rays (see Fig. 1 for light curves of utilized MAXI J1535-571 and GRS 1915+105 observations).

We have filtered our *NICER* data following standard practices, excluding South Atlantic Anomaly passages in order to identify continuous good time intervals (GTIs) which are extracted and analyzed individually. Data from detectors 14, 34, and 54 have been excised owing to a propensity for elevated noise or spurious events in

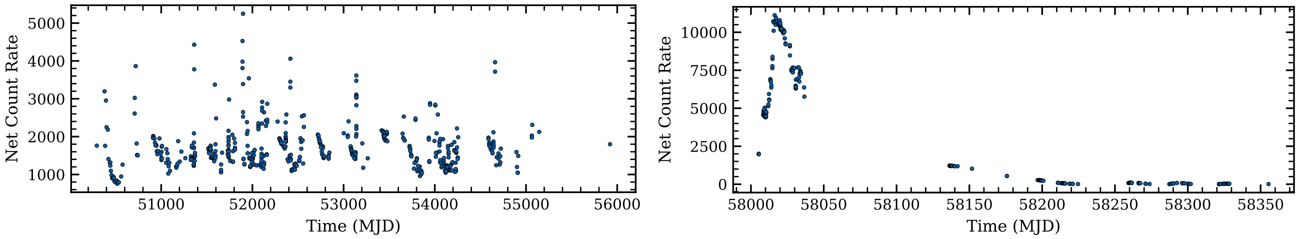

Figure 1. Light curves of GRS 1915+105 (left) and MAXI J1535−571 (right) for the observations used in this work. Net count rates are calculated as the sum of the background subtracted counts divided by observation time for every observation of each source. Note the persistent nature of GRS 1915+105 versus the transient flare of MAXI J1535−571 (reflaring epochs of MAXI J1535−571 are not included given the lack of QPOs detected there in previous works).

those detectors. Additionally, for each GTI, the average event rates of overshoot, undershoot, and X-ray events are compared amongst the detector ensemble, and any detector which has a median absolute deviation (MAD) >15 is also excised for that GTI[1]). All spectra have been corrected for deadtime (generally < 1 per cent). *NICER* backgrounds have been computed using the 3C50 background model (Remillard et al. 2022), as well as using a proprietary and similar background model which replaces the 3C50's 'hrej' and 'ibg' indexing with cutoff-rigidity 'COR_Sax' and overshoot-rate indexing. We have removed any data with a background count rate ≥5 counts/s, exclude observations for which the source-to-background count ratio is <10, and reject observations with exposure times $t \simeq 60$ s. Additionally, we require at least 5000 net source counts to ensure reliable energy and PDS results, and we consider the remaining data sufficiently bright and insensitive to the selection between these similar background models. Energy spectra have been rebinned from the 10 eV PI channels by a factor ranging from 2 to 6 in order to oversample *NICER*'s energy resolution by a factor $\gtrsim$ 2, while also requiring a minimum of 5 counts per bin. From 1 to 4096 Hz, PDS are computed using events in the energy range from 0.2 to 12 keV, for a light-curve sampling at 2^{-13} s (≈122 μs). PDS are computed individually and averaged together using 4s segments for $t < 160$ s and 16 s segments for $t \geq 160$ s. Below 1 Hz, PDS are computed by averaging together results for 128 s segments for $t \geq 128$ s 64 s segments for $64 \leq t < 128$ s and 4 s segments for $t < 64$ s. The resulting PDS is then logarithmically rebinned in ∼3% frequency intervals, the Poisson noise subtracted, and the rms^2 Hz^{-1} normalization adopted.

Although we have less MAXI J1535−571 observations with QPOs for analysis (in large part due to the source's transient nature), one benefit of using *NICER* over *RXTE* data for this source (if we could have used *RXTE* data) is that *NICER* spectral channels do not suffer from gain drift over epochs like *RXTE* PCA (which affected energy-channel conversions), and thus we can use the *NICER* energy spectra as raw inputs to our regression and classifier models, in addition to the engineered features discussed in Sections 3 and 4.2.

Overall, we selected these two sources for this initial evaluation of our methodology because they represent two very different types of LMXRBs. On one hand, GRS 1915+105 has long been known as a markedly unusual source in terms of its outburst behaviours and states (e.g. its very abnormal, three-decade long transient outburst, regular/irregular bursts, dips, etc., behaviors influenced by GRS 1915+105's orbital period and accretion disc size, the longest and largest respectively known among LMXRBs), wheres on the other hand, MAXI J1535−571 is, in comparison to GRS 1915+105, a far more typical source in terms of outburst states, QPO-spectral parameter associations, and tracks through the hardness–intensity diagram (Taam, Chen & Swank 1996; Truss & Done 2006; Nakahira et al. 2018; Bhargava et al. 2019; Cúneo et al. 2020; Koljonen & Hovatta 2021; García et al. 2022a). Hence, between these two sources we aim to evaluate our methods across a spectrum of typical to challenging spectral-timing relationships. Furthermore, in choosing objects observed with different instruments, we aim to take advantage of the different strengths of each instrument, such as the plethora of *RXTE*' observations and the high spectral resolution of *NICER* (Gendreau et al. 2012).

3 DATA ANALYSIS

3.1 Energy spectra

As previously mentioned and discussed in more detail in Section 4.2, we base our detection of QPOs on energy spectra and processed features from the energy spectra. Thus, to generate the processed spectral features we fit the energy spectra for both sources with `XSPEC` version 12.12.0 (Arnaud, Dorman & Gordon 1999) using the three-component model `tbabs*(discbb + nthcomp)`, which represents a Tuebingen–Boulder absorbed multitemperature blackbody and thermally Comptonized continuum (Mitsuda et al. 1984; Zdziarski, Johnson & Magdziarz 1996; Kubota et al. 1998; Życki, Done & Smith 1999). We fixed the equivalent hydrogen column densities to canonical values of 6×10^{22} atoms cm^{-2} for GRS 1915+105 and 3.2×10^{22} atoms cm^{-2} for MAXI J1535−571 based on Sreehari et al. (2020) and Cúneo et al. (2020), respectively, with solar abundances in accordance with Wilms, Allen & McCray (2000) and Verner et al. (1996) cross-sections (e.g. Fig. 2). We tied the `nthcomp` seed photon temperature to $T_{\rm in}$ of `discbb` for both sources, and let high energy rollover (electron temperature) freely vary between 4 and 40 keV for GRS 1915+105 and 4–250 keV during fitting for MAXI J1535−571, basing these ranges on Zhang et al. (2022) and Dong et al. (2022), respectively. For GRS 1915+105, we ignore channels <2.5 keV or >25 keV during fitting, calculate net count rate from the resulting range, and compute hardness as the sum of the ratio of the background subtracted channel net count rates for the ranges in Zhang et al. (2022), except as a proportion rather than a ratio, i.e. $\frac{[13-60]\ \mathrm{keV}}{[2-7]+[13-60]\ \mathrm{keV}}$ (see Fig. 12 for pairplot comparing spectral and timing properites of GRS 1915+105). Regarding MAXI J1535−571, we note the presence of instrumental residuals in the 1.7–2.3 keV *NICER* range, likely related to *NICER*'s Au mirror coating and residual in the Si K α fluorescence peak, and following Miller et al. (2018), we address

[1] The MAD is a robust statistic that is insensitive to outliers. 15 MAD corresponds to approximately 10σ for a Gaussian-distribution.

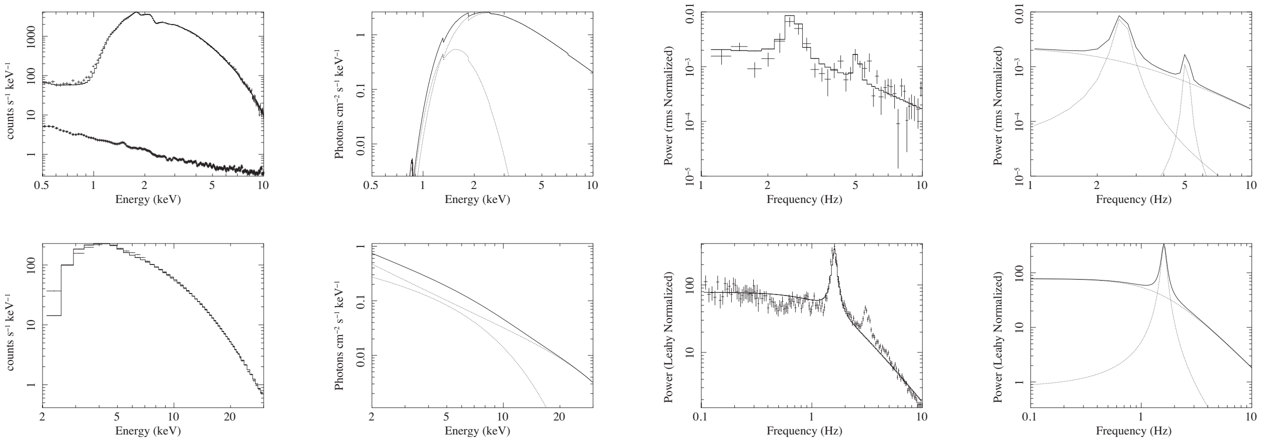

Figure 2. Example energy and power density spectra and models for MAXI J1535 observation 1050360105–21 on the top and the same for GRS 1915+105 observation 40116-01-01-07 on the bottom. For each row, from left to right, the first plot shows the energy spectrum and folded `tbabs*(nthcomp + discbb)` model, the second shows energy spectrum model alone, the third shows the power density spectrum in the relevant frequency range, and the fourth shows the best-fitting Lorentzian PDS model alone. Best-fitting QPO features have been superimposed over zero centered Lorentzians used to model the power-density continuum. Only the fundamental (i.e. first harmonic) is fit for the GRS 1915+105 QPO (as discussed in Section 3, this was an intentional choice to see how the models fair with seemingly simpler outputs).

these by excluding the 1.7–2.3 keV energy band from the spectral fitting process, and otherwise fit the range 0.5–10.0 keV. We compute net count rate normalized to the number of *NICER* detectors, and hardness ratios for MAXI J1535−571 observations as the proportion of the total net count rate contributed by the 3.0–10.0 keV range, i.e. $\frac{[3.0-10.0]\ \mathrm{keV}}{[0.5-1.7]+[2.3-3.0]+[3.0-10.0]\ \mathrm{keV}}$. Altogether, for both sources we use the net count rate, hardness ratio, asymptotic power-law photon index, `nthcomp` normalization, inner disc temperature, and `discbb` normalization for input parameters, which we discuss in more detail in Section 4.2.

3.2 Power density spectra

Throughout this work, all QPOs for both sources are parametrized as Lorentzian distributions given by equation (1),

$$A(f) = \frac{K(\frac{\sigma}{2\pi})}{(f - f_0)^2 + (\frac{\sigma}{2})^2} \qquad (1)$$

where f is frequency in Hertz, σ is full width at half maximum (FWHM), and K is the normalization, as per Arnaud et al. (1999), as shown in Fig. 3. In the case of GRS 1915+105, QPO properties are obtained by fits to PDS following Zhang et al. (2020). A QPO is considered significant when the ratio of the QPO power integral divided by its 1σ error >3 or quality factor $Q = \frac{\nu_0}{\sigma}$) >2 (Nowak, Wilms & Dove 1999), provided their frequency does not change significantly in an observation. Our primary use for this GRS 1915+105 data is to train machine learning regression models to predict the properties of the fundamental QPO feature, since *only* data with matching QPO detections are used in our GRS 1915+105 machine-learning analysis. In all, this corresponds to 554 QPOs. In contrast to this approach of fitting individual QPOs solely for regression, we use the energy and timing data from MAXI J1535−571 to explore both classification of observations into binary states of QPO presence/absence as well as multiclass QPO cardinality states[2] based on binned raw energy spectra and processed features. Additionally, for MAXI J1535−571 we predict the properties for both the fundamental and frequently appearing harmonic in the PDS based on binned energy spectra and spectral parametrizations derived from energy spectra. Our QPO detection method for MAXI J1535−571 is slightly different than that of GRS 1915+105. Specifically, we determine the presence and properties of QPOs in PDS from MAXI J1535−571 by first fitting two zero-centred Lorentzian functions to PDS and then iteratively fitting a third Lorentzian over a logarithmically sampled set of 268 frequencies f between 1 and 20 Hz, where width is kept $\sigma < \frac{f}{10}$ for an initial fit, and then freed for a subsequent refined fitting step. A peak of qualifying distance ($\Delta\chi^2$ distance to neighboring samples) and threshold (horizontal distance between samples) is identified with the `scipy` function `find_peaks` (Pedregosa et al. 2011) in the resulting distribution of $-1 \cdot \chi^2$ fit-statistic with peak height greater than the Δ10 Akaike Information Criterion (Akaike 1998). Finally, a visual inspection is required to accept a QPO candidate detection (to avoid potential spurious detections, e.g. at the frequency boundary). In 68 of observations the fundamental is accompanied by the second harmonic (the fundamental itself is called the first harmonic), in 14 observations it is alone, and in 188 observations no QPO is detected.

4 MACHINE LEARNING METHODS

4.1 Model selection

In machine learning, models can be broadly divided by two sets of classification: (i) whether they operate in a supervised or unsupervised manner; and (ii) whether they are built for classification or regression (Bruce & Bruce 2017). Since we are providing our models with explicit targets for loss minimization, our approach falls under the umbrella of supervised learning (Singh, Thakur & Sharma 2016), and as we are attempting to connect spectral information about XRBs with real-valued output vectors that describe QPOs in their power-density spectra, we also fall under (multi-output) regression (Xu et al. 2019). In selecting our machine learning models for regression, we seek those that natively support multi-output regression, incorporate capabilities for mitigating overfitting,

[2] Also called multinomial classification (Bouveyron et al. 2019), when number of classes totals to ≥3.

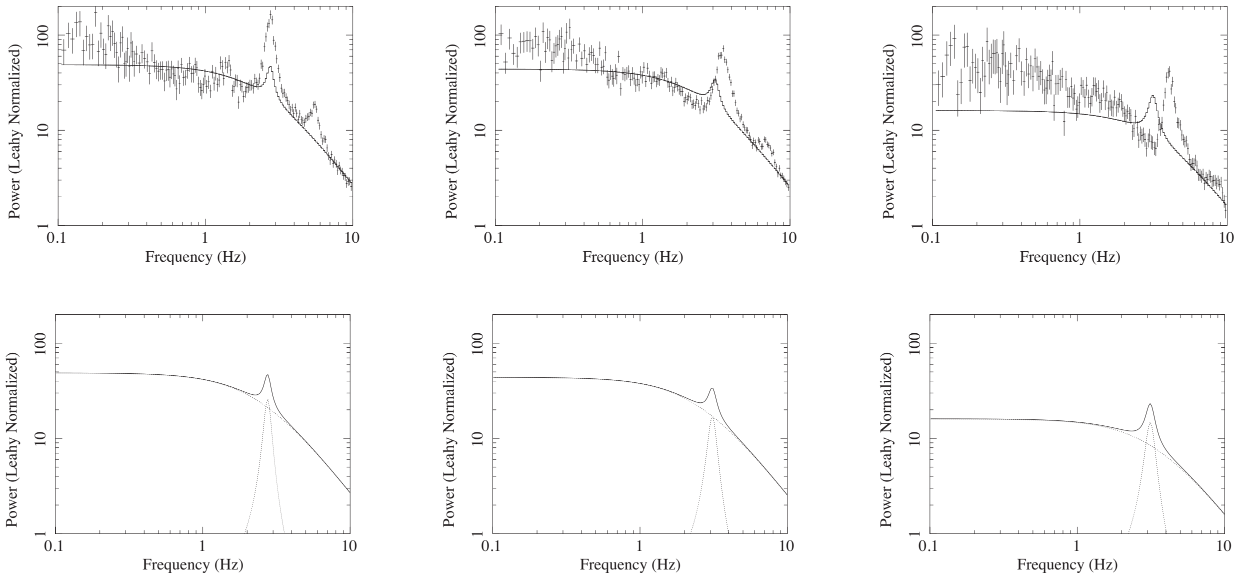

Figure 3. Example PDS with over plotted QPO predictions for the GRS 1915+105 observations 80701-01-54-02, 50703-01-28-01, 50703-01-24-01 ordered by column left to right from least (best-fitting) median to greatest (worst) Pythagorean sum of normalized errors on the three predicted QPO Lorentzian parameters (with corresponding models alone in bottom row). Note that the seemingly diminished height of the predicted QPOs is actually a consequence of how they were determined in the processing procedure, and in the case of the best observation 80701-01-54-02, the amplitude only differs by less than 0.3 per cent from the 'true' amplitude value it was predicting, as the derived amplitudes had reduced amplitudes originally.

have precedents of working successfully with medium to small sized data sets, and natively communicate feature importances. Additionally, we seek to evaluate a collection of models against each other in light of the No-Free-Lunch-Theorem (Wolpert 2002; Lones 2021).

Based on these criteria, we settle on a set of tree-based models and their descendants, specifically decision trees (Breiman 1984), random forests (Breiman 2001), and extremely randomized trees (Geurts, Ernst & Wehenkel 2006). Here, we provide a brief summary of these models for context. Decision trees are the original tree-based regression model which operate by inferring discriminative splits in data and making predictions via a series of 'if-then-else' decisions (Breiman 1984). Random forests are more powerful derivatives of decision trees, and are based on an ensemble of decision trees trained via bootstrap aggregation (Breiman 1996, 2001). By incorporating predictions from such an ensemble, random forests reduce prediction variance while increasing overall accuracy when compared to a single decision tree (Lakshminarayanan 2016). Finally, extremely randomized trees (also known as extra trees) are similar to random forests in this respect but operate with more randomization during the training process, as instead of employing the most discriminative thresholds within feature spaces for splits, extra trees select the best-performing randomly drawn thresholds for splitting rules (Geurts et al. 2006; Pedregosa et al. 2011). Details on training and optimization are given in Section 4.3, where we also discuss our steps to avoid overfitting (Bruce & Bruce 2017).

Together, these represent some of the most powerful yet lightweight machine learning models available, and meet our criteria for multi-output regression (Xu et al. 2019), robustness to overfitting (Boinee, Angelis & Foresti 2008; Ampomah, Qin & Nyame 2020), success with small/medium sized data sets (Floares et al. 2017), and feature importances (Yasodhara et al. 2021). An additional benefit of these models is that they are natively supported by the `TreeExplainer` method in the `SHAP` Python package (Lundberg & Lee 2017), which frees us from common pitfalls related to impurity and permutation based feature importances, which we discuss in more detail in Section 6. Overall, we explore all the above models in addition to ordinary linear regression (to provide a base performance comparison) for the regression cases, but focus on random forest and logistic regression (Berkson 1944) for classification cases.

4.2 Feature engineering

As Casari & Zheng (2018) detail, feature engineering is the process of transforming raw data to maximize predictive performance. After experimenting with different formats, we settled on the following in order to use derived features from spectral fits or raw spectral data as predictors and timing features as outcomes. We will hereafter refer to and experiment with two types of input data for our models: the first are rebinned net energy spectra, which we discuss below and will simply call 'energy spectra'. The second type is the combination of `XSPEC` model-fit parameters and spectrum derived features like net count rate and hardness which we will designate the 'engineered features' input type. When using engineered features for inputs, we format our input data as a matrix composed of vectors containing the net count rate, hardness ratio, asymptotic power-law photon index, `nthcomp` normalization, inner-disc temperature, and `discbb` normalization for every observation. Hereafter, we refer to and present these values by the letters $\{A, B, C, D, E, F, G\}$ as shorthand. This input structure is visualized in equation (2) as follows:

$$\mathrm{IN}_{m\times 7} = \begin{bmatrix} A_1 & B_1 & C_1 & D_1 & E_1 & F_1 & G_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_m & B_m & C_m & D_m & E_m & F_m & G_m \end{bmatrix} \tag{2}$$

where m is the number of observations. This format can be extended to any n-dimensional number of features, which we take advantage of when using raw energy spectra as input data. For the case of

MAXI J1535−571, we compare the predictive performance of the models and provide different insights by using raw spectral data in the form of count rate values from 19 channels, 0.5 keV wide apiece spanning the energy range [0.5–10.0] directly as the input vectors within the input matrix, similar to Pattnaik et al. (2020). This coarse spectral input strikes a balance between sparsity and precision, allowing us to determine importances for specific 0.5 keV ranges while not overwhelming the models with too many input features given the overall sample size (Raudys & Jain 1991; van de Schoot & Miočević 2020). With regards to regression, our QPO output matrix is similarly formatted as a vector matrix, with rows that match by index to vectors in the input matrix, but with an important addition regarding ordering (detailed below). A significant challenge relates to the prediction of not only the presence versus absence of QPOs in a given PDS, as well as (for present cases) the specific number of QPOs and the physical parameters of each QPO present. Over the course of an outburst, the number of QPOs present can change, as these are transient phenomena (Remillard et al. 2006; Ingram & Motta 2019). We account for this challenge of variable output cardinality by first identifying all QPO occurrences associated with an observation. Then, we order these occurrences and their features in a vector of length $L = N_f \times \max(N_s)$, where N_f is the number of features describing every QPO (e.g. $N_f = 3$ for frequency, width, and amplitude), and N_s is the maximum number of simultaneous QPOs observed in any particular PDS in a data set. We then structure each output vector as a repeating subset of features for every QPO contained, and order these internal QPO parametrizations by frequency. If one or more of these occurrences are not detected in a PDS, their feature spaces in the vector are populated with zeros. This allows us to circumvent the aforementioned difficulty with variable output cardinality, because the models will learn during training to associate indices populated with zeros as QPO non-detections (Chollet 2017). As with input features, equation (3) provides a visualization of the general QPO matrix output returned by our model, where each row corresponds to one observation matched with a row in the input matrix (both out of m total observations).

$$\mathrm{OUT}_{m\times n} = \begin{bmatrix} f_{1,1} & \sigma_{1,1} & K_{1,1} & \dots & f_{1,n} & \sigma_{1,n} & K_{1,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ f_{n,1} & \sigma_{m,1} & K_{m,1} & \dots & f_{m,n} & \sigma_{m,n} & K_{m,n} \end{bmatrix} \quad (3)$$

In the case of MAXI J1535−571, the maximum number of QPOs simultaneously observed in a PDS is two, and each QPO is described in terms of its frequency, width, and amplitude, so the output matrix takes the shape OUT = $m \times 6$. Since we only regress for the fundamental in the GRS 1915+105 PDS, its output matrix takes the form OUT = $m \times 3$. Prior to reformatting the data in this manner, we applied a columnar min–max standardization to the XSPEC, and hardness input features, as well as the QPO Lorentzian output features, which linearly transformed each distribution into a $[\max(x'), \min(x')] = [0.1, 1]$ range (as opposed to the traditional $[0-1]$ range given our decision to denote QPO non-detections with zero values) while preserving their shapes, according to equation (4; Kandanaarachchi et al. 2019).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \times \frac{\max(x') - \min(x')}{\min(x')}. \quad (4)$$

This step is necessary to prevent features with relatively larger absolute amplitudes receiving undue weight, and it also frees the models from dependency on measurement units (Han, Kamber & Pei 2012; Akanbi, Amiri & Fazeldehkordi 2015). We did not apply this standardization step to channel count and net count rate input features, however, as the imposition of *a priori* theoretical limits to these features is not as readily justifiable (Pattnaik et al. 2020).[3]

4.3 Training, validation, and hyperparameter tuning

To better understand our models in different data combinations and minimize statistical noise, while guaranteeing every observation gets included in a training, as well as at a separate time, validation instance, we employ a repeated k-fold cross-validation strategy (Olson & Delen 2008; Vanwinckelen & Blockeel 2012) for model evaluation (as opposed to solely using a default proportion-based train-test split). According to this procedure, our data are first split into a 90 per cent training and validation set, and then a 10 per cent held out test set. Before evaluating the models on this test set, the training and validation set is randomly split into $k = 10$ folds. Given the relative class imbalance in the MAXI J1535−571 data in favour of observations without QPOs, for MAXI J1535−571, the folds for both regression and classification cases are also stratified during splitting, which means each fold maintains the same proportion of observations with QPOs (Ma & He 2013). Then, every model is evaluated on each unique fold after being trained on the remaining folds, with the individual k-fold performance taken as the mean of these evaluations across the ten folds. We repeat this process five times (randomly shuffling the data between each iteration), and the final score for each model is calculated as the mean performance across the ten k-fold instances, either as the f-score for classification cases (a harmonic mean of the precision and recall), or the median absolute error for regression (Pedregosa et al. 2011; Kuhn & Johnson 2019). Random initialization is kept the same between models to make sure each model is trained/tested on the same data within each fold, and to ensure fair comparison between these models, each was subject to automatic and individualized hyperparameter tuning via grid search prior during this evaluation (Dangeti 2017). The specific hyperparameter values from which combinations were derived and evaluated for each model are presented in Table 1.

4.4 Feature selection

Through feature selection, it is generally important to deal with potential multicollinearity by calculating variance inflation factors (VIF) and removing features with VIF values $\gtrsim 5$ (Kline 1998; Sheather 2008). However, we have chosen not to remove potentially collinear features prior to regression for the following reasons: first, the tree based models like random forest that we focus on are by design robust from the effects of multicollinearity (Strobl et al. 2008; Chowdhury et al. 2021). Second, since multicollinearity only affects the estimated coefficients of linear models, but not their predictive ability, applying a linear model to potentially collinear data is perfectly reasonable in our case,

[3] Standardization prior to splitting data into train and validation sets does not impair our model's predictive validity when input features are derived from XSPEC because its pre-adjusted inputs will always be constrained within the theoretical bounds applied during standardization for each feature (e.g. Γ will always initially range between $x - y$ for a source, where x can be a hard lower limit like Γ = 1.1 and y can be the corresponding hard upper limit during fitting, such as Γ = 5).

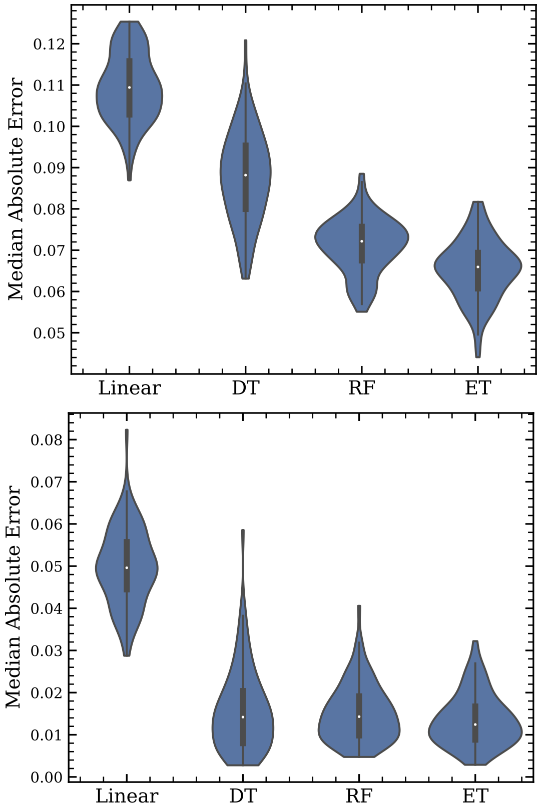

Figure 4. Gaussian kernel density estimate violin plot representations of aggregated median absolute error for each tested model across $k = 10$ validation folds repeated $r = 5$ times on GRS 1915+105 (feature input) data (top) and MAXI J1535−571 (feature input) data (bottom). The abbreviations DT, RF, and ET stand for the decision tree, random forest, and extra tree models, respectively. As further discussed in Section 5, linear regression is outperformed by the classical machine learning models models across folds for each repetition round. Furthermore, the two ensemble tree based models clearly outperform the single decision tree model, which is to be expected.

as we are using the linear model solely as a baseline against which we will compare the predictive capabilities of the more complicated random forests model; i.e. as we are applying the linear model, we are not interested in its components (Lieberman & Morris 2014; Mundfrom, Smith & Kay 2018). We will, however, revisit multicollinearity when we interpret feature importances in Section 5.

5 RESULTS

5.1 Regression

As demonstrated in Fig. 4, on average our tree-based models outperform linear regression in every regression case, regardless of source or input feature type. Interestingly, as shown in Fig. 6 and Fig. 7, linear regression also seriously struggles to correctly assign 0 values to observations lacking QPOs for both processed and rebinned energy spectra input data, a problem not faced by the other models (except random forest with rebinned energy spectra to a lesser degree). Furthermore, linear regression always has higher dispersion in the relationship between actual and predicted QPO frequency. Yet, despite their unified superiority versus linear regression, the machine learning models do differ significantly within fold amongst themselves, as shown in Fig. 4, 5, 6, and 7. Specifically, although decision tree provides a notable improvement in dispersion between true and predicted values, as well as a slope between these closer to unity, it is by far bested by random forest, and extra trees. Two additional interesting divergences in model performance occur between the sources, as well as between their input types. Regarding the former, all models trained and evaluated on GRS 1915+105 data have more overall dispersion and slopes tending further away from unity in their mapping between true and predicted frequency when compared to the same models for MAXI J1535−571 QPOs with processed input features. This can be clearly seen when comparing Fig. 5 with Fig. 6. The superior performance of the algorithms on MAXI J1535−571 are surprising for several reasons: first, with GRS 1915+105 the models never face the problem of false negatives or false positives because there are no QPO-absent data in this set. In contrast, MAXI J1535−571 observations are of varying composition, imbalanced in favor of QPO absence. Second, GRS 1915+105 has around two times more total observations, and around six times more observations with QPOs than MAXI J1535−571; in most cases training models on more data leads to corresponding increases in accuracy (Brefeld et al. 2020; Kalinin & Foster 2020). However, this assumption may not hold in instances like this, where models are being tested on different objects, as there may exist fundamentally stronger/more pronounced associations between spectral and QPO in one of the systems. The most likely reason for the inferior performance on GRS 1915+105 QPOs is that the underlying relationships between the input and output QPO features are likely more convoluted for GRS 1915+105, which is understandable given GRS 1915+105 has long been known to have complex variability states, and is in fact a bit of an oddball among black-hole systems. Additionally, potential confusion could arise because the models fitted on fundamental QPOs only in GRS 1915+105 intentionally lack the freedom to predict aspects about harmonics, which could lead to these models to potentially confuse signals for harmonics with fundamentals (this is an unexpected insight from our initial decision to only predict for the fundamental in GRS 1915+105 in an effort to explore how the models behave with simpler output space). Finally, to evaluate the performance of the multioutput aspect of the regression, we carry out pairwise non-parametric two-sided goodness-of-fit Kolmogorov–Smirnov (KS) tests on permutations of QPO parameter residual arrays (Massey 1951; KS 2008), and fail in all instances to reject the hypothesis that any pair of distributions of residual arrays between actual and predicted QPO parameters are not drawn from the same distribution ($p > 0.76$ for all GRS 1915+105 and $p > 0.99$ for all MAXI J1535−571 residual pair permutations, regardless of input type). This shows that the models do not favour any particular QPO parameter in their regression and instead regress for each with statistically insignificant differences in accuracy (i.e. accuracy is not different for QPO features, both for the fundamental, as well as the harmonic when present). As for the second interesting divergence in model performance (by input type), surprisingly there is a pronounced difference in model performance when these regression models are trained on processed features as opposed to rebinned energy spectra: in all model cases, dispersion and slope both drastically worsen when models rely on the rebinned energy spectra directly. This is shown for MAXI J1535−571 regression between Figs 6 and 7 demonstrates that although the models could hypothetically learn some lower level representation of the concepts of hardness, overall net count rate, etc. from the data and not require the engineered features, with the amount of data provided engineered features provide significant additional insight for the models to base decisions on that,

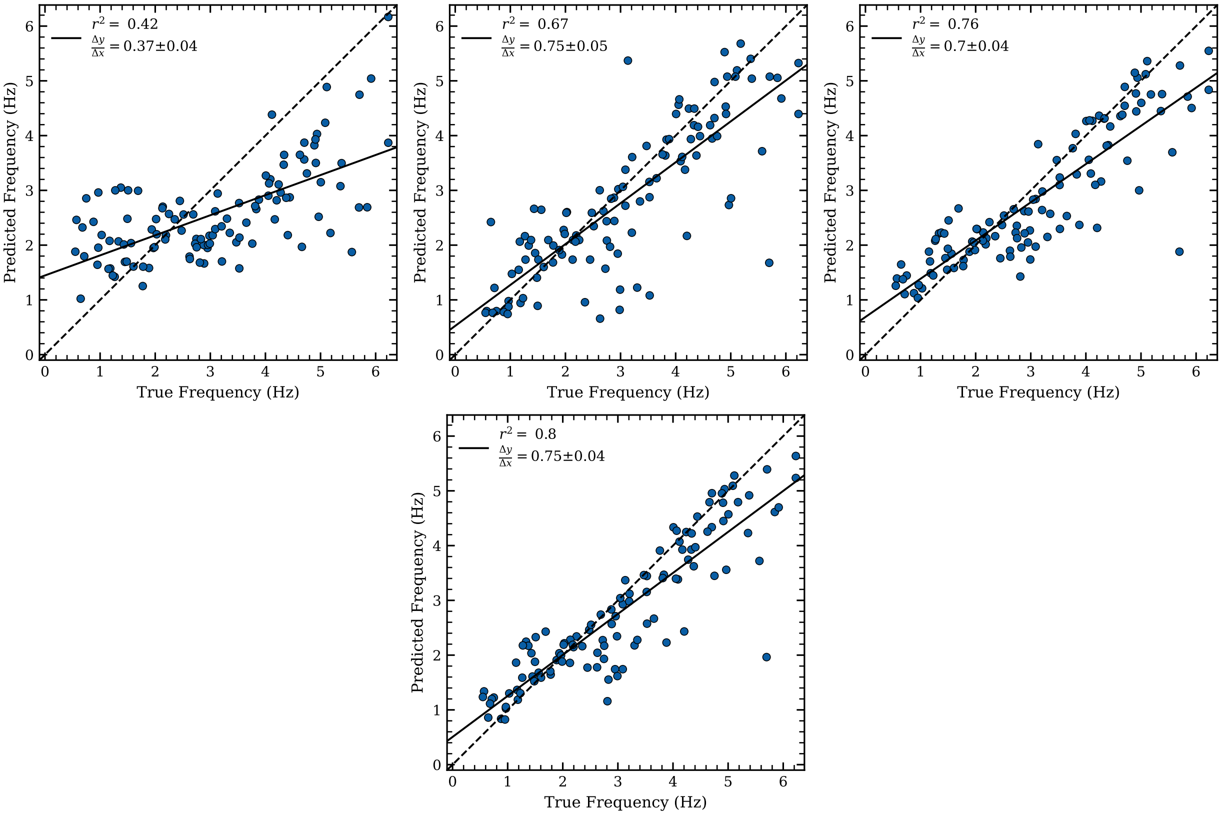

Figure 5. A results regression plot for all QPOs predicted from the test set for the source GRS 1915+105 as returned (from left to right) by linear regression, decision tree, random forest, and extra trees. The best models – random forest and extra trees – both minimize dispersion between true and predicted values (as quantified by r^2), while simultaneously producing the most 1:1 relationships between them (as quantified by best-fitting slope).

exceeding what is provided by energy spectra alone. This would be an interesting idea to investigate with deep learning methods, which would far exceed these classical models' ability to learn abstractions in the data through automated feature extraction (Nadeau & Bengio 2004).

5.2 Classification

At least for MAXI J1535−571, binary classification of QPO absence/presence appears to be a fairly trivial task, as shown by the confusion matrices of the first repetition tenth folds in Fig. 8. Additionally, as Fig. 8 also shows, our logistic regression classifier corollary to linear regression performs just as well as random forest in terms of accuracy and other classification metrics when trained on processed input data, with negligible difference for rebinned energy spectra as well. This is corroborated by the corresponding ROC curves also shown in Fig. 8. The ROC curves show how a model has optimized between specificity (on the abscissa) and recall (also known as sensitivity; on the ordinate), with the ideal model displaying an ROC curve enclosing an area under curve (AUC) of 1 (Bruce & Bruce 2017). The curves in Fig. 8 represent the average ROC and AUC values with 1 ± σ deviations across all folds and repetitions evaluated. Both logistic regression and random forest decrease in average AUC when trained on rebinned energy spectra, but the decrease is most dramatic for logistic regression. We also present multiclass classification results for multinomial logistic regression and random forest based on processed and rebinned energy spectra input data in Fig. 9. In the case of processed input data, random forest clearly outperforms logistic regression, but both models actually experience noted decreases in accuracy when tasked with predicting multiple outputs corresponding to the actual number of QPOs in a MAXI J1535−571 observation based on rebinned energy spectra input. In fact, in the case of energy spectra inputs, random forest actually performs worse than logistic regression. Overall, the decreased performance of both models here is likely do to the class imbalance in the data set (as mentioned in Section 3), which gives the models very few single QPO observations to use as training data per round.

6 DISCUSSION

Now that we have demonstrated QPOs properties can be predicted – and in the following section show how features useful to these predictions can be analysed – on the sources MAXI J1535−571 and GRS 1915+105 individually, we propose the next step would be to apply these methods in a future work on source-heterogeneous input data, a capability we intentionally incorporate into our QPOML library. To achieve this, it would be beneficial to construct a large standardized data base of QPO and spectral data with a scope *á la* Corral-Santana et al. (2016), for which the wealth of *RXTE* observations will prove invaluable. Additionally, while increasing

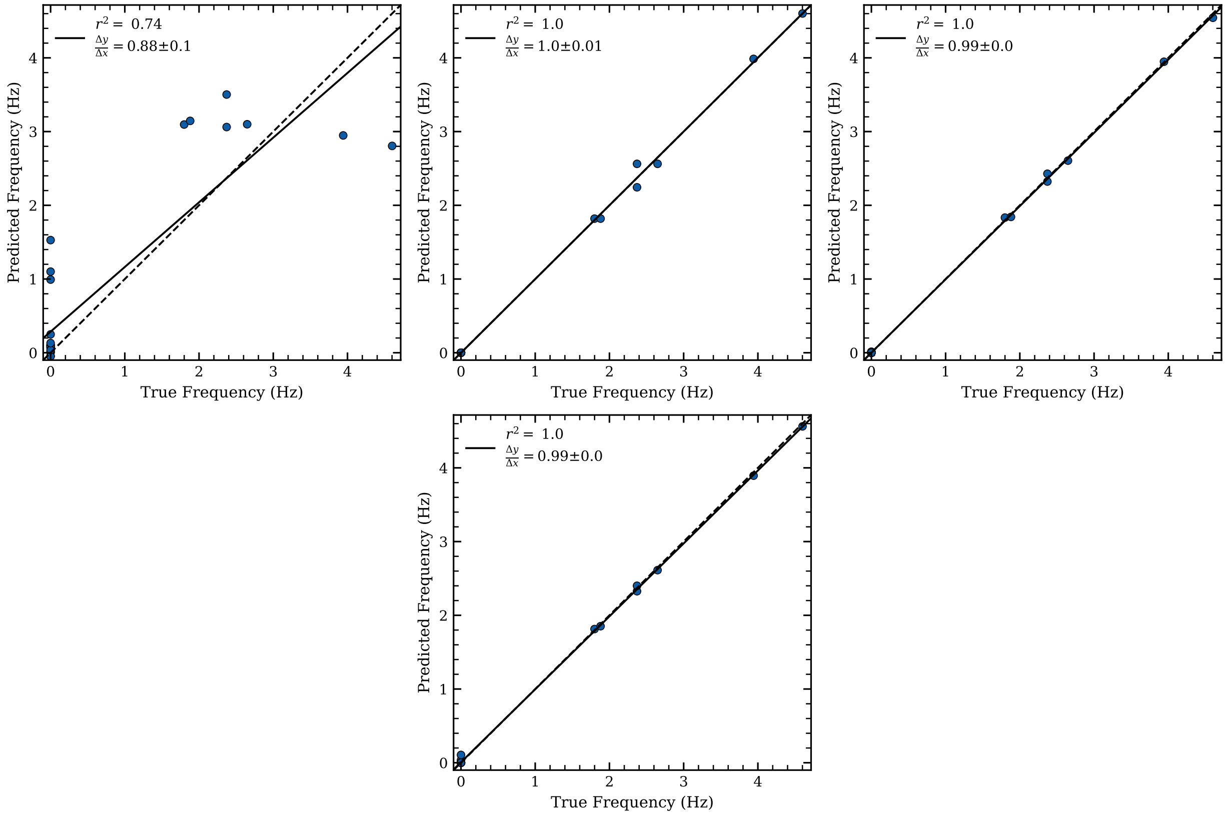

Figure 6. Same as Fig. 5, except for MAXI J1535−571 observations (processed feature input). The lesser number of points in these plots stems from both the smaller sample size of MAXI J1535−571 observations, as well as the clustering of values correctly predicted as zeros at the point (0,0) where points cannot be seen individually in this plot).

source sample size like this, it would also be fruitful to include NS LFQPOs and kHz QPOs in a follow-up study to generalize between sources, because unlike BH XRBs, NS XRBs are pre-dominantly persistent and have significantly more observations with QPOs in archival *RXTE* data in general (Méndez et al. 1999; Migliari, van der Klis & Fender 2003; Belloni, Méndez & Homan 2005; Raichur & Paul 2008). That being said, the likely trade-off of using *RXTE* data for these sources is that these QPOs will be predicted based on engineered XSPEC features instead of raw spectra given gain drift, as was the case with our analysis of GRS 1915+105 versus MAXI J1535−571. Another potential avenue for extending this work would involve exploring new input features to associate with QPOs, such as BH spin, mass, inclination, jet properties, and QPO phase lags, and tracking the importance of variable features throughout outburst and accretion states to see if they evolve in tandem. Including scattering fraction as an input parameter promises interesting results as well, because QPO frequency and scattering fraction exhibit a correlation for sources like MAXI J1535−571 but an anti-correlation for other objects including GX 339–4, H1743−322 and XTE J1650−550 (Garg, Misra & Sen 2022). Finally, how these non-parametric machine learning models interact with the polynomial/exponential versus sigmoidal relationship between frequency and power-law index for some BHs versus NSs (Titarchuk & Shaposhnikov 2005), as well as how well models trained on distinct outbursts of certain objects perform for outbursts withheld from their training, would both also be of interest if these models are applied on samples that differ not only by source, but also by source type (BH or NS). Now, we turn to discussing feature importances in Section 6.1 and statistically compare the models we used throughout this work in Section 6.2.

6.1 Feature importances and interpretation

Feature importances refer to the relative attributed weights a model gives to different input features (Saarela & Jauhiainen 2021). In other words, they are measures for how helpful different features are for the model in making correct predictions, regardless of whether these predicted values are categorical or real-valued (Fisher, Rudin & Dominici 2018). Before we discuss these, however, we will briefly describe our efforts to ensure the interpretability of our machine learning models. Interpretability is defined parsimoniously by Miller (2017) as the degree to which a human can understand the cause of a decision. Since most of our models are intrinsically complex (except for linear and logistic regression and decision trees), we seek *post hoc* interpretability through feature importances (Vieira & Digiampietri 2022). These values should not be interpreted as substitutes for other e.g. parametric importances, because they seek to explain how a machine learning model learns and interacts with its data. However, we believe that properly calculated feature importances may offer alternative helpful insight about the origins of QPOs, and we therefore take steps to avoid common pitfalls associated with these measures. For example, although it is common to discuss default impurity-based feature importances, this approach is flawed because it is both biased towards high-cardinality numerical input features, as well as computed on training set statistics, which means it may not accurately generalize to held-out data (Pedregosa et al. 2011).

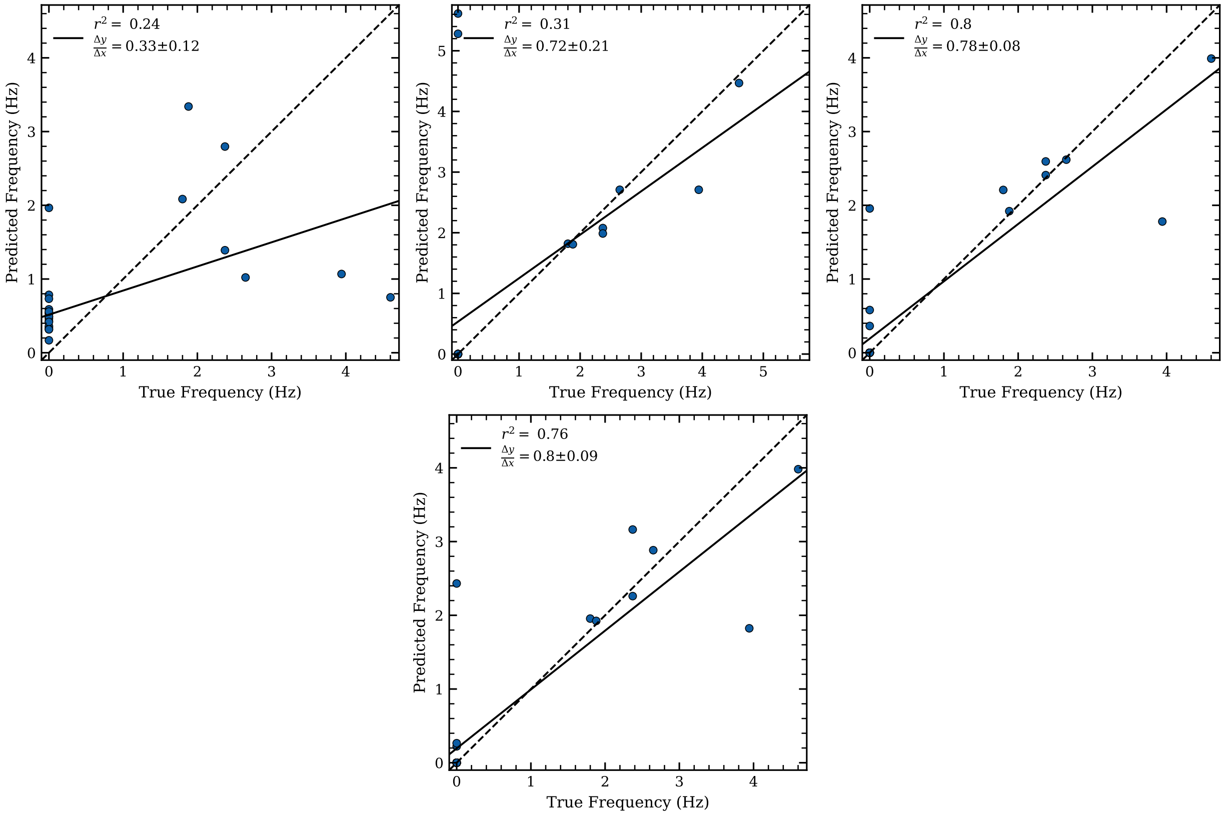

Figure 7. Same as Fig. 6, except for MAXI J1535−571 observations (rebinned energy spectra as inputs). Note the increased dispersion and much less 1:1 relationships between true and predicted values for every model in the these plots compared to their equivalents in Fig. 6.

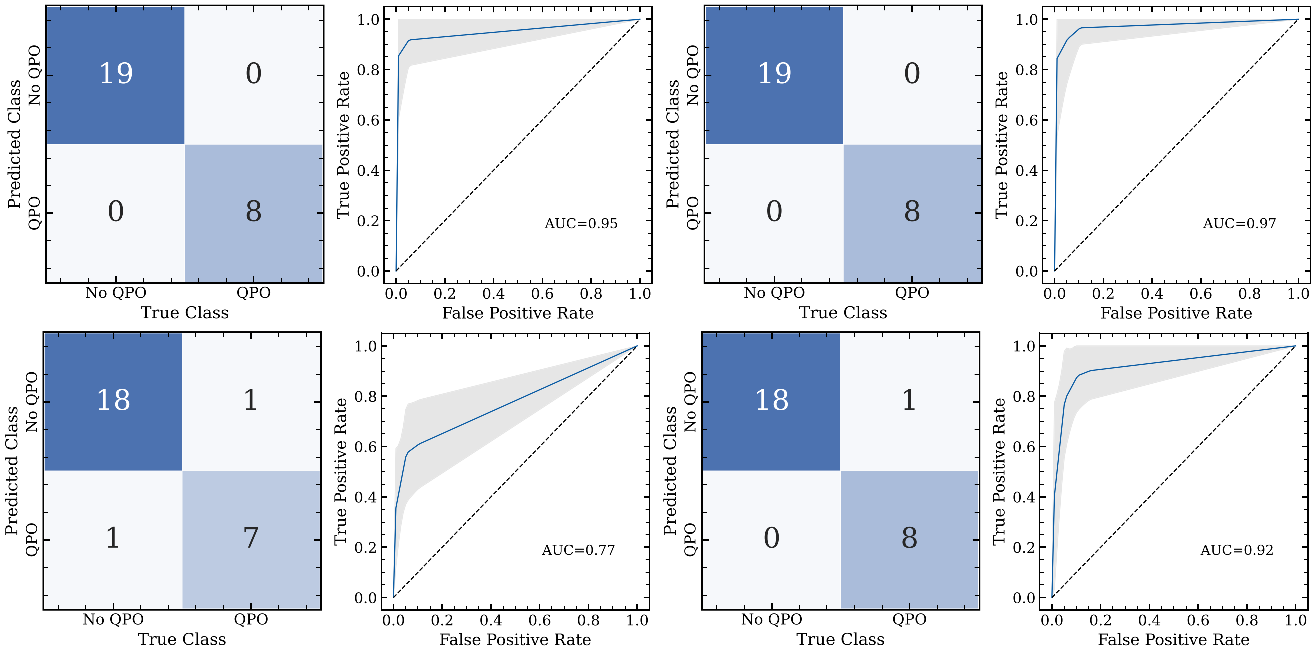

Figure 8. Confusion matrices and ROC Curves with labeled AUC values for MAXI J1535−571 binary classification cases. The left pairs correspond to logistic regression, whereas the right correspond to random forest. The confusion matrices are taken from the first tenth fold, whereas the ROC curves are averaged across all folds with ±1σ deviations denoted by the grey regions. The superior performance of the models working from processed inputs in the top row compared to their rebinned energy spectra input analogues in the bottom row is intriguing and discussed in more detail in Section 5.

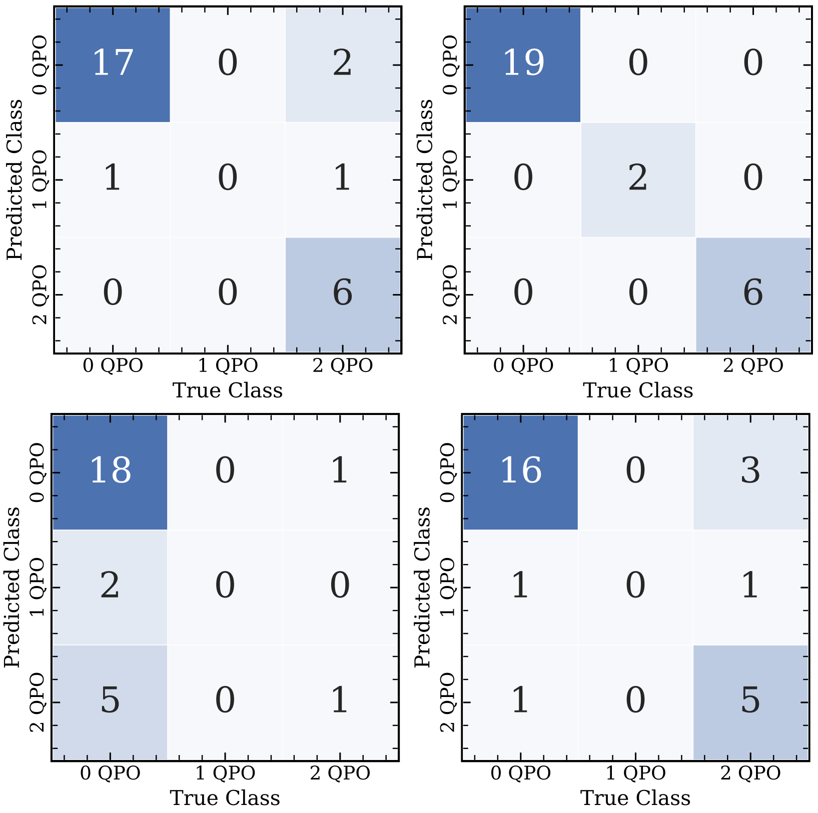

Figure 9. Confusion matrices for multiclass MAXI J1535−571 output, where the left column corresponds to logistic regression, the right column to random forest, the top row to processed input features, and the bottom row to rebinned energy spectra input features. Although only the accuracy of logistic regression decreases from binary to multinomial classification based on processed `XSPEC` input features, both models are significantly more inaccurate for the multinomial case based on energy spectra inputs compared to either binary case in Fig. 8.

Additionally, although permutation importances are commonly put forward as a superior alternative, these suffer from multicollinearity, as in the process of permutating single features, an impactful feature could be erroneously ascribed as having little-to-no effect on model performance if it has high correlation with another feature (Strobl et al. 2007; Nicodemus et al. 2010; Hooker, Mentch & Zhou 2019). Therefore, we chose to to determine feature importances with the contemporary `TreeSHAP` algorithm as implemented in the Python package `shap` by Lundberg & Lee (2017). This model extends game theoretic coalitional Shapley values to calculate SHapley Additive exPlanations (SHAP) in the presence of multicollinearity by incorporating conditional expected predictions (Shapley 1952; Lundberg & Lee 2017; Molnar 2022). As hinted earlier and detailed in Lundberg & Lee (2017) and Molnar (2022), an additional benefit of using tree based models is that through tree traversal and dynamic programming the computational cost for computing SHAP values is brought down from exponential time $\mathcal{O}(2^n)$ to $\mathcal{O}(n^2)$ polynomial time. We calculate feature importances shown in Section 5 for each model f by treating the model from the tenth fold in the first repetition as if they were taken from the test set, and averaging their $\phi_i(f, x)$ from equation (5), which represents the weighted average of differences in model performance when a feature x out of M simplified input features is present versus absent for all subsets $z'\subseteq x'$.

$$\phi_i(f, x) = \sum_{z'\subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}\left[f_x(z') - f_x(z'\setminus i)\right], \qquad (5)$$

One of the most important things shown by Figs 10 and 11is that there are significant interesting differences between the feature importances attributed to the processed features for GRS 1915+105 and MAXI 1535–571, which may be related to the nuances of the process driving QPOs in these systems. For example, in GRS 1905+105, net count rate and hardness ratio are clearly the most important features, after which importance falls precipitously and remains uniformly modest for the rest, with this proportional decrease ranging from a factor of 3 for `nthcomp` asymptotic power law to six for `nthcomp` and `discbb` normalization. Because we have used SHAP values for importance, we can rule out the un-importance of these features stemming from multicollinearity or training set artifacts, which means they could *potentially* be related to curious physical related conditions. However, there is no ambiguity about the importance of net count rate and hardness, because an XRB outburst's q-shaped state-evolution in the hardness-intensity diagrams (HIDs) is known to also be indicative of changes in timing (e.g. QPO) properties as tracked in HIDs (Motta et al. 2015; Motta 2016). This is also in agreement with the findings of fig. 2 of García et al. (2022b), in which the QPO frequency of GRS 1915+105 is shown to vary with a somewhat inverse relationship with hardness ratio across mostly horizontal and vertical gradients in inner disc temperature and power law index, respectively. In contrast to GRS 1915+105, the feature importances for both the best regression and classification models on processed MAXI J1535−571 input features favoūr a single feature above all others: `discbb` normalization (although in the case of classification, net count rate and `nthcomp`

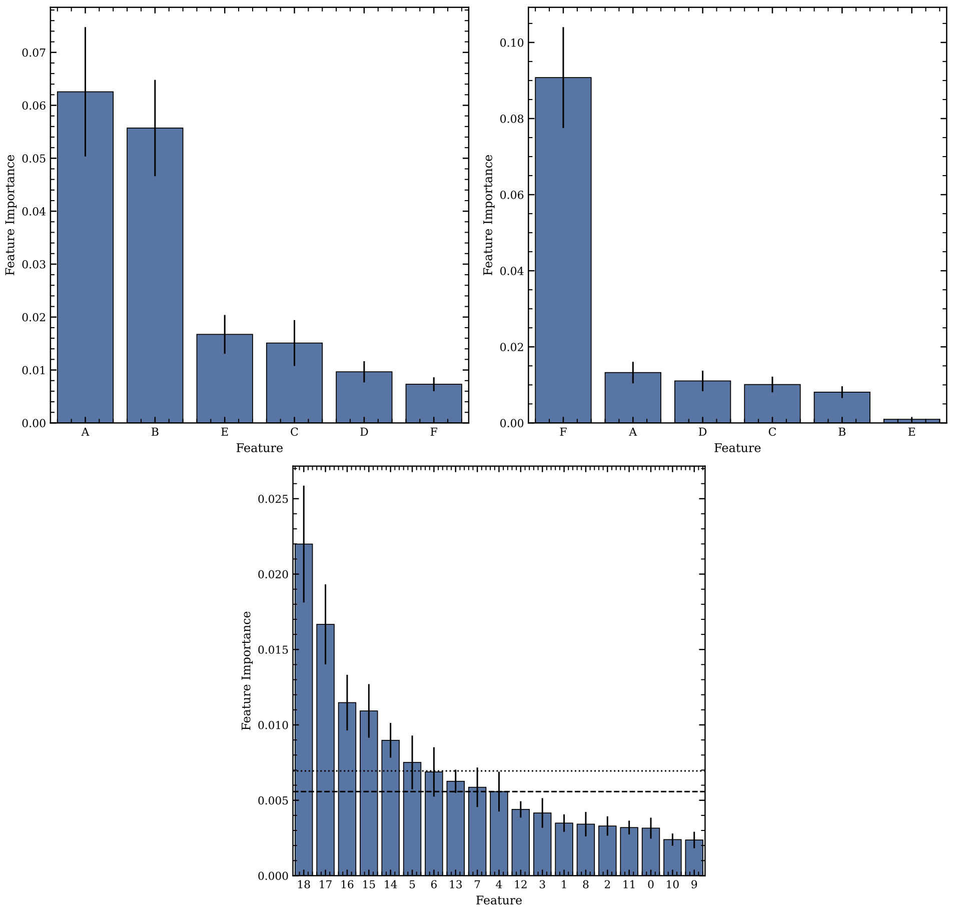

Figure 10. Tree-SHAP calculated average of absolute value SHAP feature importances for the most accurate predictive regression models for GRS 1915+105 engineered inputs (left, extra trees), MAXI J1535−571 engineered inputs (middle, extra trees), and MAXI J1535−571 energy spectra inputs (right, extra trees). The features denoted $A - F$ correspond to net count rate, hardness ratio, asymptotic power-law photon index, `nthcomp` normalization, inner-disc temperature, and `discbb` normalization features, respectively. The error bars on each importance correspond to 99 per cent confidence intervals on mean importances, the dashed line the median importance of all features, and the dotted line the mean of the same. Features corresponding to hard channel count rates are significantly more important than the median and mean feature importance, which is likely related to the higher energy origin of QPOs. An interesting difference between these plots and that for GRS 1915 + 105 in Fig. 10 is that Extra Trees primarily weights `discbb` normalization for MAXI J1535-571 regression but splits primary importance for GRS 1915+105 between the net count rate and hardness ratio engineered inputs.

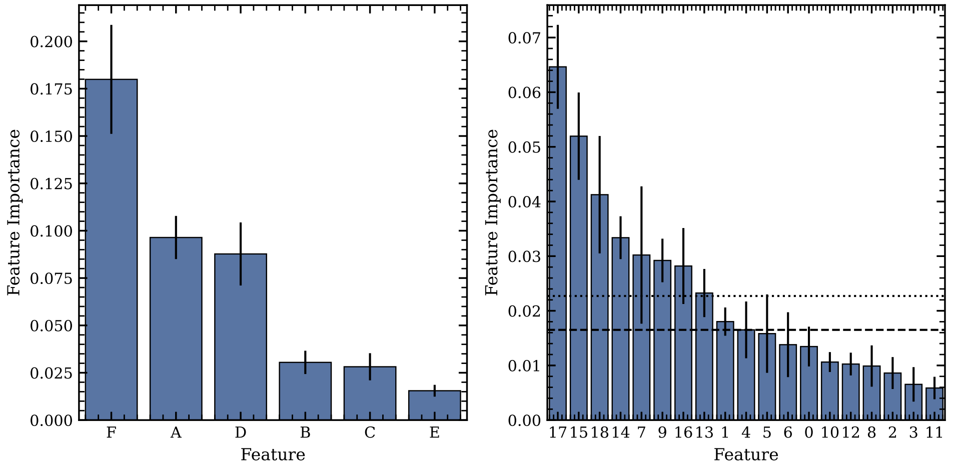

Figure 11. Similar to Fig. 10, except for the best classification models for MAXI J1535−571 binary output based on engineered inputs (left, random forest), and energy spectral inputs (right, random forest). As seen for regression, hard energy channels similarly dominant feature importances for energy spectra input, yet, while `discbb` normalization is still the most important processed feature for classification, more importance is attached here to net count rate and `nthcomp` normalization here than for regression on MAXI J1535−571.

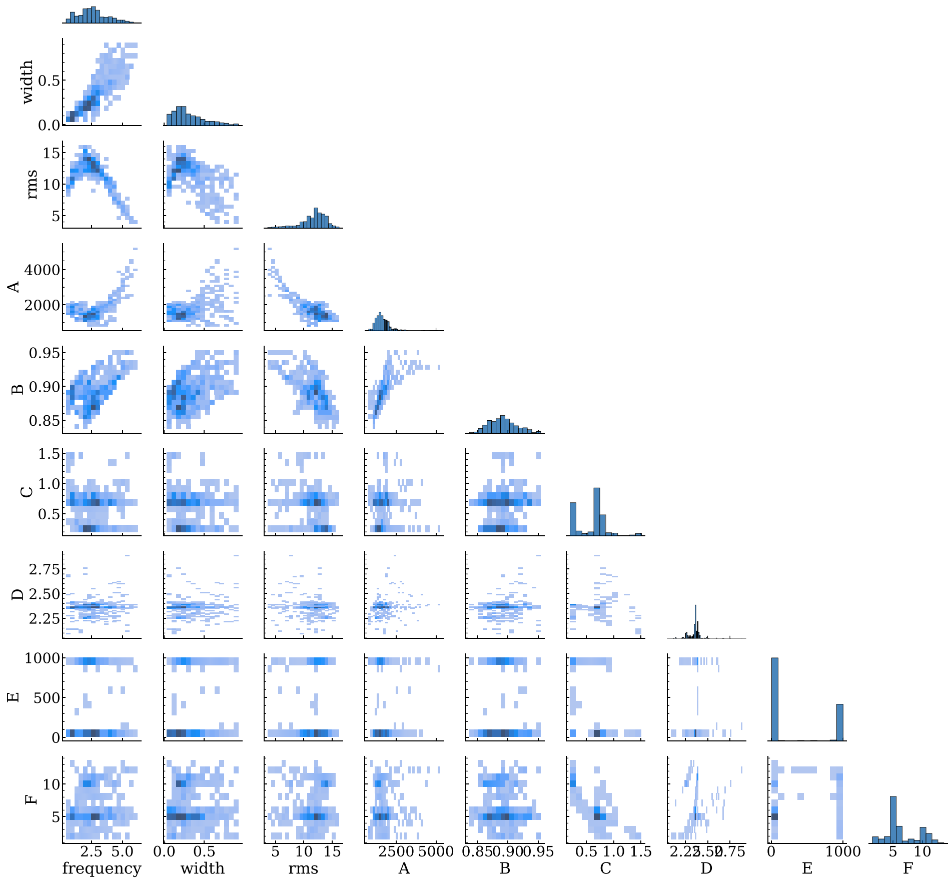

Figure 12. A pairplot displaying the pairwise relationships between engineered input and Lorentzian QPO output paramters for all GRS 1915+105 data. The letters in $A - F$ correspond to the net count rate, hardness ratio, asymptotic power-law photon index, `nthcomp` normalization, inner-disc temperature, and `discbb` normalization features, respectively.

Table 1. Feature spaces for model hyperparameter tuning.

	Decision tree	Random forest	Extra trees
`min_samples_leaf`	{1,3}	{1,3}	{1,3}
`min_samples_split`	{2,4,6,8}	{2,4,6,8}	{2,4,6,8}
`n_estimators`		{50,100,150, 200,250,500}	{50,100,150, 200,250,500}
`warm_start`		{`True`,`False`}	

normalization are still significant for MAXI J1535−571). This quantity (ignoring relativistic and plasma corrections) approximately corresponds to the projected area of the inner-disc on the sky: $N_{\rm disk} = (\frac{R_{in}}{D_{10}})^2 \cos(\theta)$, where $R_{\rm in}$ is the apparent inner disc radius in km, D_{10} is the distance to the source in 10 kpc units, and θ the angle of the disc (Arnaud et al. 1999). This prominent importance is intriguing because it implies a dependence between QPO presence and frequency on `discbb` normalization and therefore inner disc radius. This is corroborated by Garg et al. (2022), who find that QPO frequency correlates significantly with the inner disc radius for MAXI J1535−571 in data provided by *AstroSat* according to the power law relationship $\upsilon_{\rm QPO} \propto \dot{M} R_{in}^p$, where $\dot{M}$ is mass-accretion rate (Rao, Singh & Bhattacharya 2016). However, (Garg et al. 2022) do not find a clear relationship between `discbb` normalization and QPO frequencies in the ∼1.6–2.8 Hz range. Overall, the similarity in feature importances for engineered features for regression and classification in MAXI J1535−571 shows that the same features that are important in determining the parametrizations of QPOs are those important in determining their presence versus absence. Regarding the feature importances derived from the energy spectra, the highest energy channels are the most important for both regression and classification, with the five most important channel counts rates for each coming from the equivalent $[9.5-10)$, $[9.0-9.5)$, $[8.5-9.0)$, $[8.0-8.5)$, and $[7.5-8.0)$ keV channels for regression and $[9.0-9.5)$, $[9.5-10.0)$, $[8.5-9.0)$, $[8.0-8.5)$, and $[3.0-3.5)$ keV channels for classification. Notably, for both classification and regression only hard channels ≥3 keV have importances significantly greater than the mean and median importances for all features in their respective sets at the 99 per cent confidence level. The fact that the high-energy spectral data is most informative of the QPOs is interesting and we speculate that this may be related to the fact that QPOs manifest more prominently at higher energies above the disc's peak temperature. A broader perspective which generalizes these relationships to other BH systems is of high interest, but outside the scope of this work. Consequently, we are currently working on a comprehensive follow-up work, in which we will evaluate these models on data identically reprocessed for numerous BHs and NSs simultaneously. One additional difference between this preliminary work and that prospective one will be full inclusion of all LF QPO features for all sources (such as GRS 1915+105), because although focusing on the dominant frequency for QPOs in GRS 1915+105 served our purposes here, this would be a limitation in the future because such focus would not make it clear whether these trained forest methods would predict many false positives and false negatives for sources similar to GRS1915+105 that do include QBO-absent data, yet perform well none the less.

6.2 Statistical model comparison

As mentioned in Section 4, we included an ordinary least-squares model as a benchmark for their utilization. As Fig. 4, 5, 6, and 7 demonstrate, each of our models outperform linear regression. In order to assess the significance of the improvements, we employ the Nadeau & Bengio (2004) formulation of the frequentist Diebold-Mariano corrected paired *t*-test (Diebold & Mariano 1995),

$$t = \frac{\frac{1}{k\cdot r}\sum_{i=1}^{k}\sum_{j=1}^{r} x_{ij}}{\sqrt{(\frac{1}{k\cdot r} + \frac{n_{\rm test}}{n_{\rm train}})\hat{\sigma}^2}}, \tag{6}$$

where $k = 10$ and represents the number of k-fold validation folds, $r = 10$ and equals the number of times we repeated the *k*-fold procedure, x is the performance difference between two models, and $\hat{\sigma}^2$ represents the variance of these differences (Pedregosa et al. 2011). It is necessary to correct the $t-$ values in this manner because the performances of the models are correlated with each fold upon which they are tested, as some folds may make it harder for one of, or all of, the models to generalize, whereas others make it easier, and thus the collective performance of the models varies. The results of these pairwise tests for all permutations of two models on both sources is shown in Table A1.

We additionally implement the Bayesian Benavoli et al. (2016) approach, which allows us to calculate the *probability* that a given model is better than another, using the Student distribution formulated in equation (7):

$$\mathrm{St}(\mu; n-1, \bar{x}, (\frac{1}{n} + \frac{n_{\rm test}}{n_{\rm train}})\hat{\sigma}^2) \tag{7}$$

where n is the total number of samples, $\bar{x}$ is the mean score difference, and $\hat{\sigma}^2$ is the Nadeau & Bengio (2004) corrected variance in differences (Pedregosa et al. 2011). Both sets of these pairwise tests are also shown in Table A1.

Based on these tests, it is clear that extra trees significantly out performs all other models, and interestingly, that each model that follows it in decreasing order of performance is significantly better than the remaining models following it, confirming the findings in Fig. 3. In fact, in all cases of regression, the order of model performances is extra trees, random forest, decision tree, and finally, linear regression. This result is expected, with decision trees being more accurate than linear regression (because the former can leverage non-linear relationships between input features and QPOs), as well as for random forest to outperform individual decision trees (because random forests are ensemble aggregations of decision tree forests). The similar yet superior performance of extra trees in comparison to random forest is notable but not striking (Mathew 2022), yet this improvement should be considered with the additional size of an extra trees model compared to a trained random forest counterpart (this difference ranges from larger in terms of leaf count; Geurts et al. 2006). Nevertheless, based on these findings it is clear that these classical machine learning models have been able to fairly accurately optimize for individual sources. However, although extra trees may perform best in these individual source scenarios, it remains yet to be seen whether these classical models will be generalizable for accurate cross-source analyses (as proposed earlier) or if other models like neural networks will be required (Neyshabur et al. 2017). Although it may seem reasonable to combine data from these two sources and evaluate the predictive performance of these models in such a source-heterogeneous space, this would not be appropriate because any the resultant feature importances would not communicate whether or not the input engineered or raw spectral features are being leveraged for intuition into the physical state of the objects, or if their importances just reflect the models picking up on artifacts from the data generation procedure. In other words, this could be considered a form of data leakage, considering differing instrumental sensitivities, QPOs identification methods for each source, etc. (Hannun, Guo & van der Maaten 2021; Yang et al. 2022a). Hence, this provides additional motivation for follow-up, in which energy and timing spectra from a single instrument are reprocessed in an identical manner for multiple objects to prevent instrumental artifacts from contaminating the findings potentially recoverable from such a source-heterogeneous data set.

7 CONCLUSION

In this paper, we have advanced novel approaches utilizing machine learning algorithms to link energy spectral properties (as both rebinned raw energy spectra and alternatively via engineered features derived from spectral fits) with the presence and properties of QPOs prominent in power-density spectra of two low-mass XRB BH systems. Specifically, we tested a selection of tree-based classical machine learning models using engineered features derived from energy spectra to predict QPO properties for fundamental QPOs in the BH GRS 1915+105, and such derived features as well as raw rebinned energy spectra to characterize fundamental and harmonic QPOs in the BH MAXI J1535−571. Additionally, we trained classification algorithms on the same data to predict the presence/absence of QPOs, as well as the multiclass QPO state of MAXI J1535−571 observations. We compared the performance of the machine learning models against each other, and found extra trees to perform best in all regression situations for both sources. Additionally, we compared every model against simplistic linear (regression) and logistic (classification) models as well, finding the machine learning models outperformed their linear counterpart in all regression cases, with linear regression notably struggling to correctly identify observations lacking QPOs. The main findings from this study are:

(i) All tested regression models yielded significantly better results on MAXI 1535–571 versus GRS 1915+105 data, despite the latter having 6× more data with QPOs and no issue with QPO absent observations. We attributed this to the multitude of unusual variability classes unique to GRS 1915+105 among Huppenkothen et al. (2017b).

(ii) Kolmogorov–Smirnov tests on permutations of QPO parameter residuals showed that the best-fitting regression model, Extra Trees, does not favour any particular QPO parameter and instead predicts for all with equal accuracy, including those for harmonics.

(iii) Using rebinned raw spectral data as opposed to `XSPEC` derived features resulted in significantly worse performance for regression, binary classification, and multiclass classification on MAXI J1535−571 observations.

(iv) To enhance computational efficiency and ensure importance credibility, we calculated `TreeShap` feature importances immune to multicollinearity and found that for processed input features, extra trees determined the most significant features for GRS 1915+105 to be net count rate and hardness ratio, whereas the same model predicting for MAXI J1535−571 found `discbb` normalization most important, which suggests a dependence on physical inner disc radius in this case.

(v) We found almost all the rebinned channels which are the most important in determining the parameterizations of QPOs in regression are also those that are most important in determining their presence versus absence in classifying MAXI J1535−571 energy spectral data. Furthermore, for energy spectra, we found hard channels are the most important for both regression and classification, which aligns with the understanding of higher energy QPO manifestation above peak disc temperatures

(vi) We have proposed future applications of these methods that range from extending the input feature space they are tested on (e.g. scattering fraction and inclination) to moving from single source to source/source-type heterogeneous samples to achieve our original goal of inter-object generalizations since in this paper we have introduced and laid the foundation for these methods on individual objects.

Finally, we based our work on our `QPOML` Python library, from input and output matrix construction and pre-processing, to hyperparameter tuning, model evaluation, and plot generation, which were all conveniently streamlined for application and both (i) executed as ‘under-the-hood’ as possible while remaining user accessible; and (ii) easily extendable to any number of QPOs and any number of scalar observation features for any number of observations from any number of sources. This library is available on GitHub.

ACKNOWLEDGEMENTS

MM acknowledges support from the research program Athena with project number 184.034.002, which is (partly) financed by the Dutch Research Council (NUD). We also thank Virginia A. Cuneo for a helpful conversation early in this work, Michael Corcoran and Craig Gordon for assistance with some early technical issues. Finally, we thank Travis Austen with help recovering a significant amount of our work from a damaged virtual machine disc, and Brandon Barrios for Windows Subsystem for Linux advice. This work was made possible by the Neutron Star Interior Composition Explorer (NICER) and Rossi X-ray Timing Explorer (RXTE) missions, as well as data from the High Energy Astrophysics Science Archive Research Center (HEARSARC) and NASA’s Astrophysics Data System Bibliographic Services (NASA ADS). This work has been advised by AstroAI.

8 DATA AVAILABILITY

The data used for MAXI J1535−571 are available at the *NICER* archive (https://heasarc.gsfc.nasa.gov/docs/nicer/nicer_archive.html), and those for GRS 1915+105 belong to their corresponding authors and are available at the following references Zhang et al. (2020, 2022). The software used for energy spectral data analysis can be accessed from the HEASARC website (https://heasarc.gsfc.nasa.gov/lheasoft/download.html). The `QPOML` code repository can be accesed via GitHub.

Facilities:NICER, *RXTE*

*Software:Software:*ASTROPY (Astropy Collaboration 2013, 2018), KERAS (Chollet et al. 2015), MATPLOTLIB (Hunter 2007), NUMPY (Harris et al. 2020), PANDAS (Wes McKinney 2010), SCIENCEPLOTS (Garrett 2021), SCIPY (Virtanen et al. 2020), SCIKIT-LEARN (Pedregosa et al. 2011), and SEABORN (Waskom 2021).

REFERENCES

Akaike H., 1998, Information Theory and an Extension of the Maximum Likelihood Principle. Springer, New York, p. 199

Akanbi O. A., Amiri I. S., Fazeldehkordi E., 2015, A Machine-Learning Approach to Phishing Detection and Defense. Elsevier, Amsterdam, p. 45

Ampomah E. K., Qin Z., Nyame G., 2020, Information, 11, 332

Arnason R. M., Barmby P., Vulic N., 2020, MNRAS, 492, 5075

Arnaud K., Dorman B., Gordon C., 1999, Astrophysics Source Code Library, recordascl:9910.005

Astropy Collaboration, 2013, A&A, 558, A33

Astropy Collaboration, 2018, AJ, 156, 123

Belloni T., Klein-Wolt M., Méndez M., van der Klis M., van Paradijs J., 2000, A&A, 355, 271

Belloni T., Psaltis D., van der Klis M., 2002, ApJ, 572, 392

Belloni T., Méndez M., Homan J., 2005, A&A, 437, 209

Belloni T. M., Zhang L., Kylafis N. D., Reig P., Altamirano D., 2020, MNRAS, 496, 4366

Benavoli A., Corani G., Demsar J., Zaffalon M., 2016, preprint (arXiv:1606.04316)

Berkson J., 1944, J. Am. Stat. Assoc., 39, 357

Bhargava Y., Belloni T., Bhattacharya D., Misra R., 2019, MNRAS, 488, 720

Bildsten L., 1998, in Buccheri R., van Paradijs J., Alpar A.eds, NATO Advanced Study Institute (ASI) Series C, Vol. 515, The Many Faces of Neutron Stars. Kemer, Turkey, p.419

Boinee P., Angelis A. D., Foresti G. L., 2008, Int. J. Comput. Inform. Eng., 2, 2246

Bouveyron C., Celeux G., Murphy T., Raftery A., 2019, Model-Based Clustering and Classification for Data Science: With Applications in R. Cambridge Univ. Press, Cambridge

Brefeld U., Davis J., Van Haaren J., Zimmermann A., 2020, Machine Learning and Data Mining for Sports Analytics: 7th International Workshop, MLSA 2020. Springer International Publishing, Ghent Belgium

Breiman L., 1984, Classification and Regression Trees. Wadsworth International Group, Routledge, New York

Breiman L., 1996, Mach. Learn., 24, 123

Breiman L., 2001, Mach. Learn., 45, 5

Bruce P., Bruce A., 2017, Practical Statistics for Data Scientists: 50 Essential Concepts. O’Reilly Media, Sebastopol

Casari A., Zheng A., 2018, Feature Engineering for Machine Learning: Principles and techniques for Data Scientists. O’Reilly Media, Inc., Sebastopol, p. 218

Casella P., Belloni T., Stella L., 2005, ApJ, 629, 403

Castro Segura N. et al., 2022, Nature, 603, 52

Castro-Tirado A. J., Brandt S., Lund N., 1992, IAU Circ. No. 5590, #2

Chollet F., 2017, Deep Learning with Python. Manning, Shelter Island

Chollet F., et al., 2015, Keras, https://keras.io

Chowdhury S., Lin Y., Liaw B., Kerby L., 2021, preprint (arXiv:2111.02513)

Corral-Santana J. M., Casares J., Muñoz-Darias T., Bauer F. E., Martínez-Pais I. G., Russell D. M., 2016, A&A, 587, A61

Cúneo V. A. et al., 2020, MNRAS, 496, 1001

Dangeti P., 2017, Statistics for Machine Learning. Packt Publishing, Birmingham

de Beurs Z. L., Islam N., Gopalan G., Vrtilek S. D., 2022, ApJ, 933, 116

Diebold F. X., Mariano R. S., 1995, J. Bus. Econ. Stat., 13, 253

Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441

Done C., Gierliński M., 2004, Prog. Theor. Phys. Suppl., 155, 9

Dong Y., Liu Z., Tuo Y., Steiner J. F., Ge M., García J. A., Cao X., 2022, MNRAS, 514, 1422

Fabian A. C., Rees M. J., Stella L., White N. E., 1989, MNRAS, 238, 729

Fisher A., Rudin C., Dominici F., 2018, preprint (arXiv:1801.01489)

Floares A., Ferisgan M., Onita D., Ciuparu A., Calin G., Manolache F., 2017, Int. J. Oncol. Cancer Ther., 2, 13

Fragile P. C., Straub O., Blaes O., 2016, MNRAS, 461, 1356

Fudenberg D., Liang A., 2020, SIGecom Exch., 18, 4

Galeev A. A., Rosner R., Vaiana G. S., 1979, ApJ, 229, 318

Gallo E., Fender R., Kaiser C., 2005, in Burderi L., Antonelli L. A., D'Antona F., di Salvo T., Israel G. L., Piersanti L., Tornambè A., Straniero O.eds, AIP Conf. Ser. Vol. 797, Interacting Binaries: Accretion, Evolution, and Outcomes. Am. Inst. Phys., New York, p. 189

Gao H. Q. et al., 2017, MNRAS, 466, 564

García F., Méndez M., Karpouzas K., Belloni T., Zhang L., Altamirano D., 2021, MNRAS, 501, 3173

García F., Karpouzas K., Méndez M., Zhang L., Zhang Y., Belloni T., Altamirano D., 2022a, MNRAS, 513, 4196

García F., Karpouzas K., Méndez M., Zhang L., Zhang Y., Belloni T., Altamirano D., 2022b, MNRAS, 513, 4196

Gardenier D. W., Uttley P., 2018, MNRAS, 481, 3761

Garg A., Misra R., Sen S., 2022, MNRAS, 514, 3285

Garrett J. D., 2021 SciencePlots

Gendreau K. C., Arzoumanian Z., Okajima T., 2012, in Takahashi T., Murray S. S., den Herder J.-W. A. , eds, Proc. SPIE Conf. Ser. Vol. 8443, Space Telescopes and Instrumentation 2012: Ultraviolet to Gamma Ray. SPIE, Bellingham, p. 844313

Geurts P., Ernst D., Wehenkel L., 2006, Mach. Learn., 63, 3

Giacconi R., Gursky H., Paolini F. R., Rossi B. B., 1962, Phys. Rev. Lett., 9, 439

Gilmore G., 2004, Science, 304, 1915

Goodfellow I., Bengio Y., Courville A., 2016, Deep Learning. MIT Press, Cambridge

Greiner J., 2003, in van den Heuvel E. P., Kaper L., Rol E., Wijers R. A. M. J.eds, ASP Conf. Ser. Vol. 308, From X-ray Binaries to Gamma-Ray Bursts: Jan van Paradijs Memorial Symposium. Astron. Soc. Pac., San Francisco, p. 111

Greiner J., Cuby J. G., McCaughrean M. J., Castro-Tirado A. J., Mennickent R. E., 2001, A&A, 373, L37

Han J., Kamber M., Pei J., 2012, in Data Mining. Elsevier, Ghent, 83

Hannikainen D. C. et al., 2005, A&A, 435, 995

Hannun A. Y., Guo C., van der Maaten L., 2021, Association for Uncertainty in Artificial Intelligence, Conference on Uncertainty in Artificial Intelligence

Harris C. R. et al., 2020, Nature, 585, 357

Homan J., Belloni T., 2005, Ap&SS, 300, 107

Hooker G., Mentch L., Zhou S., 2019, preprint(arXiv:1905.03151)

Hunter J. D., 2007, Comput. Sci. Eng., 9, 90

Huppenkothen D., Heil L. M., Hogg D. W., Mueller A., 2017b, MNRAS, 466, 2364

Ingram A., Motta S. E., 2019, New Astron. Rev.

Ingram A., Done C., Fragile P. C., 2009, MNRAS, 397, L101

Ivezić Ž., Connolly A. J., VanderPlas J. T., Gray A., 2014, Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data, Princeton Univ. Press, Princeton

Jonker P. G., van der Klis M., Wijnands R., 1999, ApJ, 511, L41

Kolmogorov–Smirnov Test, 2008, The Concise Encyclopedia of Statistics. Springer, New York, p. 283

Kalinin S., Foster I., 2020, Handbook On Big Data And Machine Learning In The Physical Sciences (In 2 Volumes). World Scientific Series On Emerging Technologies, World Scientific Publishing Company, Singapore

Kandanaarachchi S., Muñoz M. A., Hyndman R. J., Smith-Miles K., 2019, Data Min. Knowl. Discovery, 34, 309

Kato S., 2005, PASJ, 57, L17

Kato S., Fukue J., 1980, PASJ, 32, 377

Kline R., 1998, Principles and Practice of Structural Equation Modeling. Methodology in the Social Sciences, Guilford Publications, New York

Kojima T. et al., 2020, ApJ, 898, 142

Koljonen K. I. I., Hovatta T., 2021, A&A, 647, A173

Kremer J., Stensbo-Smidt K., Gieseke F., Pedersen K., Igel C., 2017, IEEE Intell. Syst., 32, 16

Kubota A., Tanaka Y., Makishima K., Ueda Y., Dotani T., Inoue H., Yamaoka K., 1998, PASJ, 50, 667

Kuhn M., Johnson K., 2019, Applied Predictive Modeling. Springer, New York

Lakshminarayanan B., 2016, PhD thesis, UCL (University College London)

Leahy D. A., Elsner R. F., Weisskopf M. C., 1983, ApJ, 272, 256

Li X., Zheng Y., Wang X., Wang L., 2020, ApJ, 891, 10

Lieberman M., Morris J., 2014, 40, 5

Liu Q. Z., van Paradijs J., van den Heuvel E. P. J., 2007, A&A, 469, 807

Liu Q., Liu H., Bambi C., Ji L., 2022, MNRAS, 512, 2082

Lones M. A., 2021, preprint(arXiv:2108.02497)

Lundberg S., Lee S.-I., 2017, preprint(arXiv:1705.07874)

Ma Y., He H., 2013, Imbalanced Learning: Foundations, Algorithms, and Applications. Wiley, Hoboken

Massey F. J., 1951, J. Am. Stat. Assoc., 46, 68

Mathew T. E., 2022, J. Theor. Appl. Inform. Technol., 100

McClintock J. E., Remillard R. A., 2006, in, Compact stellar X-ray sources, Vol. 39. Cambridge Univ. Press, Cambridge, p. 157

McKinney W., 2010, in van der Walt S., Millman J.eds, Proceedings of the 9th Python in Science Conference. p. 56 Python for Scientific Computing Conference, Austin

Méndez M., Belloni T. M., 2021, in Belloni T. M., Méndez M., Zhang C.eds, Astrophysics and Space Science Library, Vol. 461, Timing Neutron Stars: Pulsations, Oscillations and Explosions. Springer-Verlag, Berlin, p. 263

Méndez M., van der Klis M., Ford E. C., Wijnands R., van Paradijs J., 1999, ApJ, 511, L49

Méndez M., Altamirano D., Belloni T., Sanna A., 2013, MNRAS, 435, 2132

Méndez M., Karpouzas K., García F., Zhang L., Zhang Y., Belloni T. M., Altamirano D., 2022, Nat. Astron., 6, 577

Migliari S., van der Klis M., Fender R. P., 2003, MNRAS, 345, L35

Miller T., 2017, preprint(arXiv:1706.07269)

Miller J. M. et al., 2018, ApJ, 860, L28

Mirabel I. F., Rodríguez L. F., 1994, Nature, 371, 46

Mitsuda K. et al., 1984, PASJ, 36, 741

Molnar C., 2022, Interpretable Machine Learning, 2 edn, Leanpub, Victoria

Molteni D., Sponholz H., Chakrabarti S. K., 1996, ApJ, 457, 805

Motta S. E., 2016, Astron. Nachr., 337, 398

Motta S., Muñoz-Darias T., Casella P., Belloni T., Homan J., 2011, MNRAS, 418, 2292

Motta S. E., Casella P., Henze M., Muñoz-Darias T., Sanna A., Fender R., Belloni T., 2015, MNRAS, 447, 2059

Mundfrom D., Smith M., Kay L., 2018, Gen. Linear Model J., 44, 24

Nadeau C., Bengio Y., 2004, Machine Learning, 52, 239

Nakahira S. et al., 2018, PASJ, 70, 95

Negoro H. et al., 2017a, Astron. Telegram, 10699, 1

Negoro H. et al., 2017b, Astron. Telegram, 10708, 1

Neilsen J., 2013, Adv. Space Res., 52, 732

Neyshabur B., Bhojanapalli S., Mcallester D., Srebro N., 2017, in Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R., Vishwanathan S., Garnett R.,eds. Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc., Red Hook

Nicodemus K. K., Malley J. D., Strobl C., Ziegler A., 2010, BMC Bioinform., 11, 110

Nowak M. A., Wilms J., Dove J. B., 1999, ApJ, 517, 355

Olson D., Delen D., 2008, Advanced Data Mining Techniques. Springer, Berlin, Heidelberg

Orwat-Kapola J. K., Bird A. J., Hill A. B., Altamirano D., Huppenkothen D., 2022, MNRAS, 509, 1269

Parikh A. S., Russell T. D., Wijnands R., Miller-Jones J. C. A., Sivakoff G. R., Tetarenko A. J., 2019, ApJ, 878, L28

Pattnaik R., Sharma K., Alabarta K., Altamirano D., Chakraborty M., Kembhavi A., Méndez M., Orwat-Kapola J. K., 2020, MNRAS, 501, 3457

Pattnaik R., Sharma K., Alabarta K., Altamirano D., Chakraborty M., Kembhavi A., Méndez M., Orwat-Kapola J. K., 2021, MNRAS, 501, 3457

Pearson K. A., Palafox L., Griffith C. A., 2018, MNRAS, 474, 478

Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825

Raichur H., Paul B., 2008, ApJ, 685, 1109

Rao A. R., Singh K. P., Bhattacharya D., 2016, preprint(arXiv:1608.06051)

Raudys S., Jain A., 1991, IEEE Trans. Pattern Anal. Mach. Intell., 13, 252

Reid M. J., McClintock J. E., Steiner J. F., Steeghs D., Remillard R. A., Dhawan V., Narayan R., 2014, ApJ, 796, 2

Remillard R. A., McClintock J. E., Orosz J. A., Levine A. M., 2006, ApJ, 637, 1002

Remillard R. A. et al., 2022, AJ, 163, 130

Revnivtsev M., Churazov E., Gilfanov M., Sunyaev R., 2001, A&A, 372, 138

Richards J. W. et al., 2011, ApJ, 733, 10

Rodríguez J.-V., Rodríguez-Rodríguez I., Woo W. L., 2022, Data Mining and Knowledge Discovery, 12, e1476

Ross R. R., Fabian A. C., 2005, MNRAS, 358, 211

Saarela M., Jauhiainen S., 2021, SN Appl. Sci., 3, 1

Schlegel E. M., 1995, Rep. Prog. Phys., 58, 1375

Schmidt S. et al., 2021, Phys. Rev. D, 103, 043020

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

Shapley L. S., 1952, A Value for N-Person Games. RAND Corporation, Santa Monica, CA,

Sheather S. J., 2008, A modern approach to regression with R, 2009 edn. Springer Texts in Statistics, Springer, New York

Singh A., Thakur N., Sharma A., Institute of Electrical and Electronics Engineers, Manhattan, 2016, in 3rd International Conference on Computing for Sustainable Global Development (INDIACom). p. 1310

Sreehari H., Nandi A., 2021, MNRAS, 502, 1334

Sreehari H., Nandi A., Das S., Agrawal V. K., Mandal S., Ramadevi M. C., Katoch T., 2020, MNRAS, 499, 5891

Sridhar N., Bhattacharyya S., Chandra S., Antia H. M., 2019, MNRAS, 487, 4221

Stella L., Vietri M., 1998, ApJ, 492, L59

Stella L., Vietri M., 1999, Phys. Rev. Lett., 82, 17

Strobl C., Boulesteix A.-L., Zeileis A., Hothorn T., 2007, BMC Bioinform., 8

Strobl C., Boulesteix A.-L., Kneib T., Augustin T., Zeileis A., 2008, BMC Bioinform., 9

Taam R. E., Chen X., Swank J. H., 1996, in American Astronomical Society Meeting Abstracts. American Astronomical Society, Washington, p. 35.08

Tagger M., Pellat R., 1999, A&A, 349, 1003

Tauris T. M., van den Heuvel E. P. J., 2006, in, Vol. 39, Compact Stellar X-ray Sources. Cambridge University Press, Cambridge, p. 623

Thomas D., 1952, In Country Sleep: And Other Poems. James Laughlin, New York

Titarchuk L., Shaposhnikov N., 2005, ApJ, 626, 298

Truss M. R., Done C., 2006, MNRAS, 368, L25

van de Schoot R., Miočević M., 2020, Small Sample Size Solutions: A Guide for Applied Researchers and Practitioners. European Association of Methodology Series, Taylor and Francis, London

van den Eijnden J., Degenaar N., Russell T. D., Wijnands R., Miller-Jones J. C. A., Sivakoff G. R., Hernández Santisteban J. V., 2018, Nature, 562, 233

van der Klis M., 2006, Compact stellar X-ray sources. Cambridge University Press, Cambridge, p. 39

Vanwinckelen G., Blockeel H., 2012, BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning, 39

Verner D. A., Ferland G. J., Korista K. T., Yakovlev D. G., 1996, ApJ, 465, 487

Vieira C. P., Digiampietri L. A., 2022. Machine Learning Post-Hoc Interpretability: A Systematic Mapping Study. Association for Computing Machinery, New York, NY, USA,

Virtanen P. et al., 2020, Nat. Methods, 17, 261

Wang J., 2016, Int. J. Astron. Astrophys., 06, 82

Waskom M. L., 2021, J. Open Source Softw., 6, 3021

White N. E., Holt S. S., 1982, ApJ, 257, 318

Wilms J., Allen A., McCray R., 2000, ApJ, 542, 914

Wolpert D. H., 2002, The Supervised Learning No-Free-Lunch Theorems. Springer, London, p. 25

Xu D., Shi Y., Tsang I. W., Ong Y.-S., Gong C., Shen X., 2019, preprint(arXiv:1901.00248)

Yang C., Brower-Sinning R. A., Lewis G. A., Kästner C., 2022a, preprint(arXiv:2209.03345)

Yang H., Hare J., Kargaltsev O., Volkov I., Chen S., Rangelov B., 2022b, ApJ, 941, 104

Yasodhara A., Asgarian A., Huang D., Sobhani P., 2021, preprint(arXiv:2110.00086)

Zdziarski A. A., Johnson W. N., Magdziarz P., 1996, MNRAS, 283, 193

Zhang W., Jahoda K., Swank J. H., Morgan E. H., Giles A. B., 1995, ApJ, 449, 930

Zhang L. et al., 2020, MNRAS, 494, 1375

Zhang Y., Méndez M., García F., Karpouzas K., Zhang L., Liu H., Belloni T. M., Altamirano D., 2022, MNRAS, 514, 2891

Zhu W. W. et al., 2014, ApJ, 781, 117

Życki P. T., Done C., Smith D. A., 1999, MNRAS, 309, 561

APPENDIX

Table A1. Pairwise fold corrected frequentist and Bayesian statistics for all regression model comparisons discussed in Section 6, where GRS 1915+105 models are only tested on extracted (Processed) features, whereas MAXI J1535−571 models are tested on both Processed, as well as rebinned raw energy spectra features (Spectral). Abbreviations-wise, ET (Extra Trees), RF (Random Forest), and DT (Decision Tree). The *t* values represent the fold-corrected Student's *t* values of the differences of the average residual values for each model. These are accompanied by their corresponding *p* values.

Source (Input type)	First model name	Second model name	*t*	*p*	Per cent chance first better	Per cent chance second better
MAXI J1535−571 (Spectral)						
	ET	RF	0.67	0.25	74.74	25.26
	ET	DT	0.83	0.21	79.60	20.40
	ET	Linear	5.73	0.00	100.00	0.00
	RF	DT	0.18	0.43	57.12	42.88
	RF	Linear	5.17	0.00	100.00	0.00
	DT	Linear	5.66	0.00	100.00	0.00
MAXI J1535−571 (Processed)						
	ET	DT	0.40	3.47e−01	65.45	34.55
	ET	RF	0.60	2.76e−01	72.59	27.41
	ET	Linear	11.21	9.35e−12	100.00	0.00
	DT	RF	0.15	4.40e−01	56.00	44.00
	DT	Linear	8.74	1.62e−09	100.00	0.00
	RF	Linear	9.73	1.86e−10	100.00	0.00
GRS 1915+105 (Processed)						
	ET	RF	1.25	1.07e−01	89.37	10.63
	ET	DT	4.24	4.33e−05	100.00	0.00
	ET	Linear	11.20	4.19e−16	100.00	0.00
	RF	DT	3.61	3.29e−04	99.98	0.02
	RF	Linear	9.04	9.27e−13	100.00	0.00
	DT	Linear	4.51	1.75e−05	100.00	0.00

This paper has been typeset from a TEX/LATEX file prepared by the author.