BERTground: A Transformer-Based Model of Background Spectra on the ISS-Based NICER Space Telescope

Anh N. Nhu[1, 2, 3], **Abderahmen Zoghbi**[1, 2, 3]
[1]University of Maryland, College Park, MD 20742, USA
[2]CRESST II, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA
[3]HEASARC, Code 6601, NASA/GSFC, Greenbelt, MD 20771, USA
anhu@terpmail.umd.edu, abderahmen.zoghbi@nasa.gov

Abstract

The Neutron star Interior Composition Explorer (NICER) is an International Space Station (ISS)-based Space Telescope developed by NASA and devoted to the study of high-energy X-Ray sources in the universe, including but not limited to neutron stars, pulsars, and black holes in stellar systems and active galactic nuclei (AGN). One prominent problem with NICER observations is the highly variable background spectra, obscuring actual signals of astrophysical sources and negatively affecting scientific analysis of the targets. Therefore, obtaining accurate estimations of the background spectra is crucial to filter the noise and facilitate better scientific discoveries of new astronomical objects. In this paper, we propose the very first Deep Neural Network architecture to model the NICER background spectra variation using information about the spacecraft and telescope associated with each observation. In particular, we develop a BERT-based architecture with tokenizers applied to different groups of features in our tabular dataset. We also introduce an adapted Tabular Deep Residual Network architecture as the predictor following the Transformer modules in our network. We show that our model outperforms the current state-of-the-art background model developed by the NICER team in most evaluation metrics. Finally, we discuss pathways and future work for the deployment of this model on NASA's next versions of HEASARC Software packages.

Introduction

The Neutron star Interior Composition ExploreR (NICER) is NASA's X-ray timing and spectroscopy instrument that was launched on SpaceX's Falcon rocket and aboard the International Space Station (ISS) in June 2017. It has been in operation since then and serves as a critical tool facilitating fundamental research of various astronomical events. NICER's X-ray Timing Instrument (XTI) consists of 56 different X-ray concentrator optics (XRC; Okajima et al.), each capturing 0.2 - 12 keV X-ray photons emitted from an approximately 30 arcmin2 region of sky (Markwardt et al. 2023). Similar to other space-based X-ray detectors (RXTE, XXM-Newton, NuSTAR, etc.), NICER is subject to background spectra, which consist of photons not originated from the actual astrophysical source, associated with each observation. The presence of background can result in significant distortions to the source signals, negatively affecting scientific analysis of the true target's spectral variability. Therefore, the development of an accurate, reliable background model is necessary to subtract background photon counts from raw observations. Unlike detectors that have CCD detectors which allow one to simultaneously separate photons from the source from those of the background by selecting the relevant regions on the CCD, NICER does not produce images (single-pixel detector), so estimating the background spectra is a non-trivial task.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

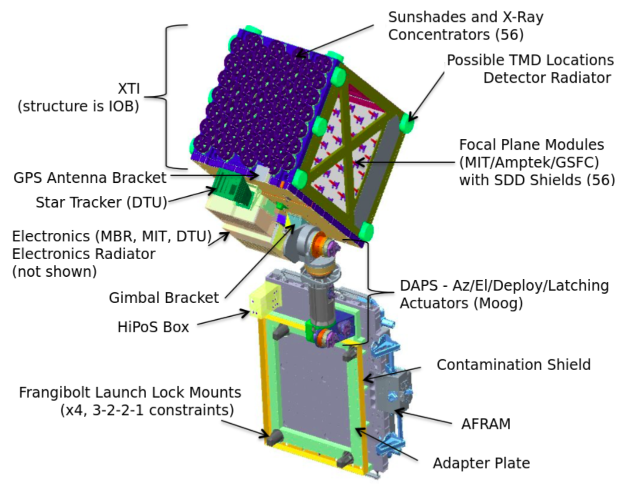

Figure 1: NICER telescope with 56 X-ray concentrator mirrors and sunshades. This figure is from NASA Goddard Space Flight Center's official NICER Mission Guide documentation (Markwardt et al. 2023).

The background spectra can originate from various sources, which can be either X-ray (e.g.: Cosmic X-ray Background) or non-X-ray (e.g.: local environment particles). Since the environment and cosmic background are highly variable depending on the spacecraft, space weather, and detector settings, background spectra can be estimated by using the information about the ISS environment and NICER detector itself. For instance, the currently deployed state-of-the-art background model, the SCORPEON (NASA 2022), estimates the background based on South Atlantic Anomaly (SAA), Cosmic Rays (COR_SAX), Polar and Precipitation Electrons, various constants (non-X-ray, cosmic X-ray background, Halo, and Local Hot Bubble). While having high overall performance, this model is relatively simple and utilizes only a few manually selected variables and priori constants. As a result, many potentially important variables, such as ISS's position, sunlight condition, detector's force trigger, and noise rate, remained unused. Furthermore, these models are either rule-based (3C50 and SpaceWeather; NASA) or non-deep-learning models (SCORPEON), which may fail to capture complex non-linear signals from the data. In fact, despite being state-of-the-art, SCORPEON only performs well for observations with backgrounds with low counts ($\leq$ 100 photons) and thus is only useful when the background is low. Given the aforementioned limitation and current advancements in Deep Learning, Deep Neural Network emerges as a promising approach to model and extract highly non-linear signals for accurate background estimation. However, there is no existing work proposing Deep Learning approaches for NICER background spectra estimation. Motivated by this gap and by the limitations of existing background models, in this paper, we propose BERTground, a novel Transformer-based Deep Neural Network architecture, for estimating background spectra in NICER observations based on 46 different parameters. Specifically, BERTground consists of two primary components: **(1)** BERT modules (Devlin et al. 2019) with categorical tokenizers and **(2)** a novel Tabular Deep Residual Network adapted from Residual Networks (ResNet; (He et al. 2016)) in Computer Vision. The proposed model outperforms deployed state-of-the-art model by a significant margin, especially for observations with excessive background, showing promises as the next generation of NICER background. Furthermore, it lays the foundation for future work on Deep Learning-based background spectra modeling.

In summary, the contributions of this paper are **(1)** the very first work on Deep Neural Networks for modeling background in NICER observations, **(2)** a novel Transformer-based architecture that outperforms current state-of-the-art and other Deep Learning models in most metrics, **(3)** a benchmark for future development of NICER background models, and **(4)** a new tool to be deployed on NASA's HEARSARC Software (HEASoft) packages for background subtraction in astrophysical research.

Related Work

3C50 (Remillard et al. 2022), Space Weather (NASA 2020d), and SCORPEON (NASA 2022) are 3 deployed background models on NASA's HEASoft that are currently being used as background filter for NICER observations. The 3C50 is a rule-based background model using two primary parameters: IBG and HREJ. In the context of this model, IBG is defined as the in-focus event rate at 15-18 keV, which is beyond the effective area of NICER's optics, while HREJ is a parameter of the particle event rate at 3-18 keV stems from the outer edges of active silicon and underneath the metal collimator. The values of these two parameters are binned into different ranges and divided into different corresponding grids (clusters). In stage 1 of the model, different observations are assigned to different grids based on their measured IBG and HREJ, and a background spectrum is matched and used as the prediction based on each observation's assigned grid. The authors also proposed to use Stage 2 of the model to further subtract soft X-ray background associated with ISS's daytime observations, which is simply done by additively combining spectra from both stages. In summary, 3C50 relies on two main parameters (IBG and HREJ), dividing their parameter space into 33 different clusters and then estimating the background by averaging the photon counts of all observations in each cluster. However, one significant drawback of 3C50 is that if an observation cannot be matched to any predetermined cluster, it is ignored and thus no estimation can be made.

An alternative to 3C50 is the Space Weather model available on HEASoft (NASA 2020d). Space Weather relies on parameters about the spacecraft environment, including local cutoff rigidity (Earth's magnetic field shielding measure) and planetary KP index (disturbance of magnetic field), to model NICER background spectra. One similarity between 3C50 and Space Weather is that they both are relatively simple rule-based models built upon observations of only 2 to 3 parameters, which limits their capability to capture complex background variations.

Contrary to 3C50 and Space Weather, SCORPEON is the very first parameterized model of NICER background. Although this model has not been formally published in conference proceedings or journals, it is considered state-of-the-art and has been employed extensively by the astrophysics community. The fundamental approach of SCORPEON is to separately estimate different background components, such as constant X-ray / non-X-ray backgrounds and Solar Wind Charge Exchange, and then linearly combine them to produce the final background prediction. One advantage of SCORPEON is that it is a physics-informed model, explicitly integrating physical phenomena as the strong priori estimates for the model. However, since the model is still limited to only 2 main variables (COR_SAX and overshoot rate), thus its capability is still not fully maximized.

NICER Background Spectra Dataset

To model the background spectra, we train our model on blank sky regions that are known to contain no astrophysical source. In other words, any X-ray or particle photon observed in those regions is purely generated by the background spectra. There are 7 different blank sky regions in total, designed by NICER as BKGD_RXTE_{1,2,3,4,5,6,8}. In this context, RXTE is the Rossi X-ray Timing Explorer (NASA 2013), a NASA satellite also dedicated to observations of high-energy X-ray sources. Sky region BKGD_RXTE_7 is excluded from our dataset since it is later discovered that this region actually contains a bright star, a soft X-ray source (Remillard et al. 2022). For this reason, BKGD_RXTE_7's spectra are not purely from the background and thus cannot be used as training targets for background models.

There are 3,037,654 total seconds of exposure from 7 different background regions, which corresponds to more than 3 million independent 1-second observations. Motivated by prior works and by the physical dependency of background variations on ISS's local environment and detector configurations, we focus on 46 different MKF parameters (NASA 2020a,b) associated with each observation as the predictors of its background spectrum. The list of MKF variables is provided in Table 1, and their detailed descriptions are available in NASA's HEASoft documentations (NASA 2020a,b,c). The targets are X-ray spectra, each comprising photon counts in 1180 different energy bands from 0.2 to 12 keV.

In summary, our dataset has a tabular format with 3,037,654 rows, 46 features, and 1180 target columns. In other words, the background model is $f : \mathbb{R}^{46} \rightarrow \mathbb{R}^{1180}$. The dataset was obtained from NICER observations available in the HEASARC database on the SciServer platform (NASA 2023a).

Methodology

Our model directly maps the predictor variables to the photon counts in all 1180 energy bins (0.2 to 12 keV) for each 1-second observation. The architecture consists of two primary novelties: (1) Group Tokenizers and (2) Tabular Deep Residual Network with dense skip connections. The Group Tokenizers summarize physically-related groups of features into tokens and feed them into BERT encoders (Devlin et al. 2019). Tabular Deep Residual Network is introduced to replace Multi-Layer Perceptron (MLP) in Transformer architecture, exploiting raw features for background estimation.

Feature Group	MKF Variables
ISS's location & pointing angle	ROLL; ANG_DIST; XTI_PNT_JITTER; SAT_LAT; SAT_LON; SAT_ALT; ATT_ANG_AZ; ATT_ANG_EL; RAM_ANGLE; EAST_ANGLE; ANG_DIST_X; ANG_DIST_Y
Solar's position	SUNSHINE; TIME_SINCE_SUNSET; SUN_ANGLE; BETA_ANGLE; LOCAL_TIME; MOON_ANGLE
Geomagnetic	SAA; SAA_TIME; COR_ASCA; COR_SAX; MCILWAIN_L; MAGFIELD; MAGFIELD_MIN; MAG_ANGLE; AP8MIN; AE8MIN; KP; SOLAR_PHI; COR_NYM; ELV; BR_EARTH
NICER's noise	FPM_RATIO_REJ_COUNT; FPM_FT_COUNT; FPM_NOISE25_COUNT; FPM_TRUMP_SEL_1500_1800; FPM_RATIO_REJ_300_1800; FPM_SLOW_LLD; MPU_NOISE20_COUNT; MPU_NOISE25_COUNT
NICER's overshoots & undershoots	FPM_OVERONLY_COUNT; FPM_UNDERONLY_COUNT; FPM_DOUBLE_COUNT; MPU_OVERONLY_COUNT; MPU_UNDERONLY_COUNT

Table 1: The list of features in each group. Variables are grouped based on their physical relatedness to each other.

Feature Group Tokenizers

Feature Group Tokenizers map a group of input features into more meaningful tokens, which are then fed into Transformer modules for contextualization of the space conditions and detector's configurations. Our design of Transformer tokenizers is based on FT-Transformer (Gorishniy et al. 2021), an adapted Transformer architecture for tabular data. However, instead of outputting a token for every single numerical feature, we group the features based on their physical relatedness and then produce a single token for each group. The benefits of group tokenizers are two-fold: first, they model the relationships between relevant variables to produce more meaningful representations; second, compared to the single-feature tokenizer in Gorishniy et al., group tokenizers allow feature summarization and reduce redundant tokens, resulting in decreased model size, training time, and hardware requirements. Another key novelty is the introduction of Sigmoid activations into the tokenization process. Specifically, instead of linearly mapping numerical features to tokens, we enforce each element in the token to take a probabilistic value between 0 and 1. The intuition is to simulate the behaviors of decision trees, where each token can be viewed as a categorical embedding of a feature group.

Formally, suppose $\boldsymbol{X}_G \in \mathbb{R}^{b\times m}$ is the group G of m features in b observations, d is the token size, then the feature group tokenizer is defined as $T : \mathbb{R}^{b\times m} \rightarrow \mathbb{R}^{b\times d}$. Accordingly, the tokens of feature group G is:

$$[token]_G = T(X_G) = \sigma(W_G \cdot X_G + b_G) \in \mathbb{R}^{N\times d} \quad (1)$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$; W_G, b_G are the linear weights and biases, respectively, of the tokenizer for feature group G. Although different groups have varied numbers of features ($\mathbb{R}^m$), their tokens share similar dimensions ($\mathbb{R}^d$). Information about feature groups is described in Table 1.

Tabular Deep Residual Network (ResNet)

In Transformer architectures, MLP is the final predictor following Transformer modules. This approach performs well for most domains, including Natural Language Processing and Computer Vision, because the raw data representations are sparse and are not meaningful by themselves. However, for tabular data, the information is dense and can be used to perform prediction directly. As a result, introducing skip connections between the input layer and hidden layers can be particularly useful to reuse such information. Furthermore, skip connections between layers can resolve the exploding / vanishing gradients problem (He et al. 2016), allowing us to construct deeper predictors with higher expressive power. Motivated by these observations and by the successes of Residual Network (ResNet) in Computer Vision (He et al. 2016), we adapt ResNet to tabular data as the tabular deep residual network. The architecture of ResNet is shown in Figure 2.

BERTground

Our proposed BERTground architecture combines Feature Group Tokenizers with the Tabular Deep Residual Network

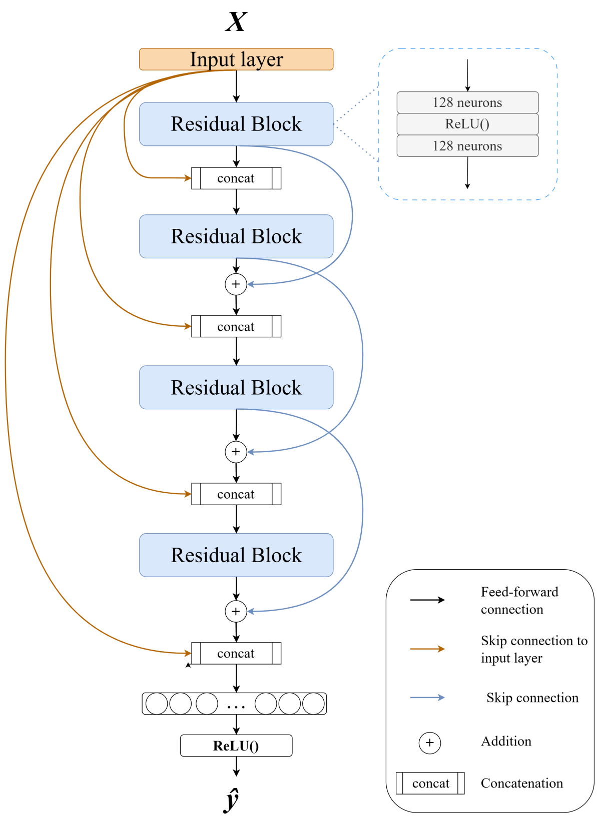

Figure 2: Tabular Deep Residual Network architecture with skip connections from the input layer. ReLU is used in the output layer since the photon counts are non-negative.

- a simple yet highly effective reconstruction of the BERT architecture (Devlin et al. 2019) specifically for NICER background estimation. The Transformer modules are the original BERT modules, which prepend the [CLS] token to the token stack. This token captures the contextual summarization for background estimation in each observation using the self-attention between different groups of physical properties explicitly defined in feature group tokenizers. Figure 3 outlines the algorithmic flow of BERTground.

Loss Function and Model Initialization

Background prediction is a multi-target problem: each spectrum consists of photon counts from 1180 energy channels (0.2 to 12 keV; 0.01 keV per channel). Therefore, the prediction accuracy depends on both the spectra intensity (total photon counts) and the spectra shape (distribution of photon counts). We employ Mean Squared Log Error (MSLE) in our objective function for model optimization. Suppose $f : \mathbb{R}^{46} \rightarrow \mathbb{R}^{1180}$ is the model, y_i is the ground-truth, and $\hat{y}_i = f(X_i) \in \mathbb{R}^{1180}$ is the estimated background spectra for NICER observation i^{th}. Accordingly, let $s_i = \sum_{e=0.2}^{12keV} y_i^e \in \mathbb{R}$ and $\hat{s}_i = \sum_{e=0.2}^{12keV} \hat{y}_i^e \in \mathbb{R}$ be the total photon counts across all energy channels for observation i^{th} in ground-truth and predicted spectra, respectively. In this context, y_i^e is the photon counts in energy channel e in observation i^{th}'s background spectrum. The loss function is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{spectra} + \lambda_2 \mathcal{L}_{intensity} \tag{2}$$

where:

$$\mathcal{L}_{spectra} = \mathcal{L}_{soft} + \mathcal{L}_{midSoft} + \mathcal{L}_{midHard} + \mathcal{L}_{hard} \tag{3}$$

and

$$\mathcal{L}_{intensity} = \frac{1}{N} \sum_{i}^{N} || \log \hat{s}_i - \log s_i ||^2 \tag{4}$$

In Eq. 3, we separate loss for each energy range to avoid optimization bias to energy level(s) with higher photon counts, implicitly optimizing spectra shape accuracy. The formal definition of the loss function for each energy level is (step interval = 0.01 keV):

$$\mathcal{L}_{soft} = \frac{1}{N \times 101} \sum_{i}^{N} \sum_{e=0.2}^{2.2keV} || \log \hat{y}_i^e - \log y_i^e ||^2$$

$$\mathcal{L}_{midSoft} = \frac{1}{N \times 129} \sum_{i}^{N} \sum_{e=2.21}^{3.5} || \log \hat{y}_i^e - \log y_i^e ||^2$$

$$\mathcal{L}_{midHard} = \frac{1}{N \times 500} \sum_{i}^{N} \sum_{e=3.51}^{8.5keV} || \log \hat{y}_i^e - \log y_i^e ||^2$$

$$\mathcal{L}_{hard} = \frac{1}{N \times 450} \sum_{i}^{N} \sum_{e=8.51}^{12keV} || \log \hat{y}_i^e - \log y_i^e ||^2$$

For the total loss in Eq. 2, we use $\lambda_1 = 1$ and $\lambda_2 = 0.05$ in our experiments. Lastly, we employ Uniform Xavier (Glorot and Bengio 2010) as the weight initializer.

Experimental Results

Train, Validation, and Test Split

In deployment, background models are used on sky regions with actual astrophysical sources, thus the model must be generalizable to sky regions that it was not trained on. To evaluate model robustness to unseen sky regions, we use sky regions 4 and 5 (RXTE_BKGD_4 and RXTE_BKGD_5) observations as validation and test set, while the training set contains samples from all remaining sky regions. This split strategy resulted in 2,665,710 (81.14%), 247,580 (7.54%), and 371,944 (11.32%) observations in train, validation, and test set, respectively.

Model Training Details

Our model was implemented with PyTorch (Paszke et al. 2019) and trained by minimizing the loss function specified in Equation 2. We used Adam optimizer (Kingma and Ba 2015) with initial learning of 10^{-3} and batch size of 512. The training batches are randomly re-shuffled after each training epoch. To avoid oscillations around local minima, we periodically decay the learning rate by a factor of 10 every 250 training epochs, helping the model to capture complex patterns from the data and to avoid overfitting to the

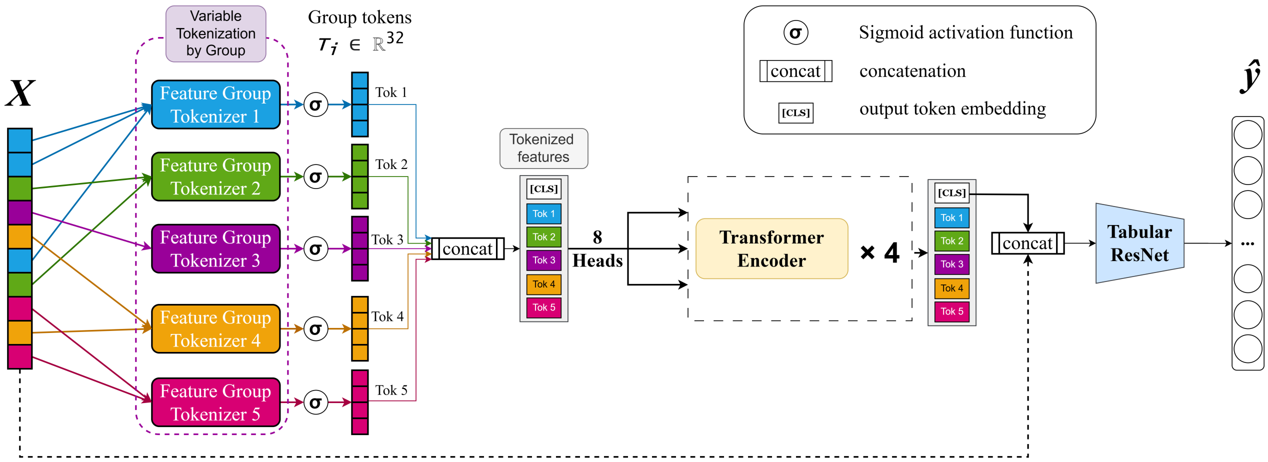

Figure 3: BERTground architecture with Feature Group Tokenizers for different feature subsets, grouped by their physical affinity (for e.g: NICER pointing conditions, Solar positions, Geomagnetic activities, X-ray noises, and telescope's conditions). 4 Transformer Encoders were used, each with 8 attention heads. Transformer Decoder blocks are not employed in BERTground.

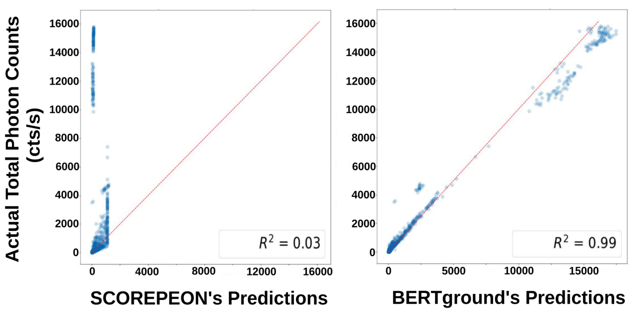

Figure 4: Predicted Background Intensity accuracy between SCORPEON (left) and BERTground (right). Background intensity is total photon counts with energy ranging from 0.2 to 12 keV per second. Most observed backgrounds have low to medium intensity ($\leq$ 100 photons), while few have extreme intensity ($\geq$ 1,000 photons).

noises (You et al. 2019). Early stopping was also used if validation loss stopped decreasing in 60 epochs. It took approximately 820 epochs and 23.4 hours to train BERTground using a single NVIDIA Tesla A100-PCIe-80GB in a Jupyter environment with 540 GB of Virtual Memory.

Results and Discussion

To evaluate the performance of our model, we qualitatively compare background predictions of BERTground and SCORPEON to the testing ground-truth spectra. First, we investigate whether the predicted background intensity is close to the actual intensity. Background intensity is formally defined as s_i and $\hat{s}_i$ in Equation 4. In simple terms, background intensity is the total photon counts aggregated over all energy channels (0.2 to 12 keV) in a particular observation. The intensity analysis illustrates whether there is background over-subtraction or under-subtraction problems when we apply the model to astrophysical observations. Over-subtraction results in distortion and information loss within the source spectra, while under-subtraction leaves background photon influx unfiltered, hindering meaningful scientific interpretation from the observation. The intensity of BERTground and SCORPEON's estimations versus the ground truths is shown in Figure 4. The scatter plots show that BERTground outperforms SCORPEON by a significant margin, yielding a R^2 score of **0.99** compared to SCORPEON's R^2 score of **0.03**. The primary problem with SCORPEON is that it fails to accurately estimate excessive backgrounds with more than 1000 photon counts. SCORPEON's accuracy on low-count backgrounds is also less accurate, as demonstrated by large deviations from the best-fit line.

Intensity alone does not suffice as a comprehensive analysis of the model performance since it does not account for the spectra shape, which is critical for many scientific analyses. Therefore, following the analysis done in (Remillard et al. 2022), we also include the analysis of Normalized Spectra of Good Time Interval (GTI). In the context of our paper, a normalized spectrum of a time interval refers to photon counts averaged along the time axis for each energy bin. The mapping function of the normalization process is $f : \mathbb{R}^{T \times E} \rightarrow \mathbb{R}^{E}$, where T is the number of time steps in the interval and E is the number of energy channels. Furthermore, Good Time Intervals (GTIs) are intervals during which observations are sufficiently high-quality, and there must be no gap of more than 5 seconds between any two consecutive observations. We also impose a restriction that each GTI must have at least 120 seconds of exposure. In total, we have 730 GTIs with an average exposure of 510.33 seconds. The comparative Normalized Spectra and Light Curves of SCORPEON and BERTground for 2 sample GTIs are shown in Figure 5. Compared to SCORPEON's predictions, both the normalized spectra shape and light curves of BERTground better fit ground-truth observations. Despite being slightly noisier, the normalized spectra of BERTground still resemble the overall ground-truth spec-

Background model	Overall (all cases)	Extreme background ($\geq$ 1000 photons)	High background (100 − 1000 photons)	Medium background (10 − 100 photons)	Low background (< 10 photons)
SCORPEON	2.9337	36.6201	0.6313	0.1585	0.0586
Linear Regression	**2.8229**	**35.2262**	0.7291	0.1765	0.0658
MLP	**2.2678**	**28.288**	**0.6313**	0.1834	**0.0548 ***
ResNet	**2.0868**	**26.0278**	**0.6247**	0.1621	**0.0552**
BERTground	**2.0061 ***	**25.0204 ***	**0.5528 ***	**0.1572 ***	**0.0582**

Table 2: Quantitative comparison between different models on the test set (sky region RXTE_BKDG 5). The performance is measured in terms of Root Mean Squared Error (RMSE). Observations are split into sub-cases based on the background intensity, which is the total photon counts in all X-ray channels (0.2 to 1.2 keV). In this table, cells in **bold** indicate better performance compared to SCORPEON model, while the * notation denotes the best-performing model.

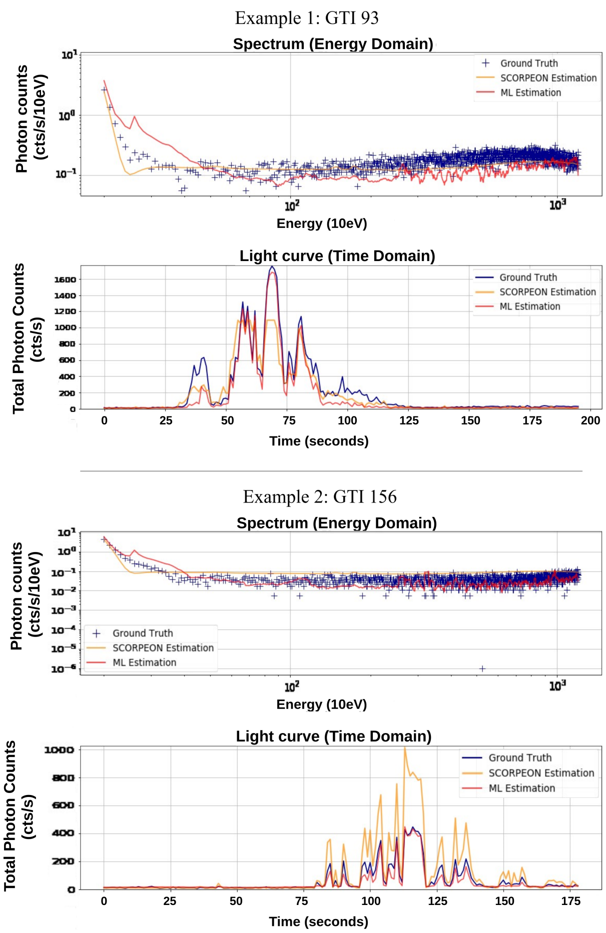

Figure 5: Comparisons: BERTground and SCORPEON. Normalized Spectra are the time-average photon counts per energy channel, while the Light Curve is the total photon counts across all channels per second (intensity).

tra shapes while fitting better into the background's spectra.

We quantitatively compare BERTground to SCORPEON and other parameterized ML models, including Linear Regression, Multi-Layer Perceptron, and ResNet on different scenarios: Overall (all observations), extreme-background, high-background, medium-background, and low-background cases. The comparative performance, measured using RMSE, is reported in Table 2. Quantitative results indicate that BERTground outperforms all other models to a significant degree in most scenarios, most notably in cases with extreme and high background intensity. Although it is not the best-performing model for low-count observations, BERTground has more competitive accuracy than SCORPEON. Finally, all Deep Learning models have lower overall prediction errors compared to SCORPEON, demonstrating the effectiveness of Deep Learning approaches for modeling NICER background spectra.

Conclusion and Future Work

In this paper, we introduce the very first Deep Neural Network-based approach as a tool to model background spectra on NICER. In particular, our proposed model, BERTground, consistently outperforms the state-of-the-art scientific model (SCORPEON) and other deep learning models (MLP and ResNet) in most scenarios, especially for observations with very high background intensity. Furthermore, this work is proof of concept that Deep Learning can efficiently estimate background spectra using MKF variables, serving as foundational work for the future development of Deep Learning-based background models. In future work, we will introduce temporal inductive bias into the model by training Transformer or LSTM-based architectures on a sequence of observations. This inductive bias is helpful to capture naturally time-dependent background components, including but not limited to emission-line components in Non-X-ray Background (NXB) spectra (Tawa et al. 2008).

Path to Deployment

BERTground shows potential as a robust background filter for analyzing high-energy X-ray sources like Neutron Stars, X-ray binaries, and supermassive black holes in AGN. We plan to deploy BERTground as a new background model in future versions of NASA's HEASoft (heasarc.gsfc.nasa.gov/docs/software/heasoft (NASA 2023b)). Dr. Abderahmen Zoghbi is an astrophysicist and HEASARC key staff at NASA Goddard Space Flight Center, leading the technical development of NASA astrophysics's cloud science platform, including HEASoft's software environment, data storage, and official documentation. Together with the NICER team, he will guide BERTground's integration into HEASoft, aligning it with astrophysicists' needs.

Acknowledgments

This research is based upon work supported by NASA under award numbers 80GSFC21M0002 and 80NSSC23K0333.

References

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W.; and Titterington, M., eds., *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, 249–256. Chia Laguna Resort, Sardinia, Italy: PMLR.

Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; and Babenko, A. 2021. Revisiting Deep Learning Models for Tabular Data. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Markwardt, C.; Gendreau, K.; Arzoumanian, Z.; Corcoran, M.; and Ray, P. 2023. NICER Mission Guide. https://heasarc.gsfc.nasa.gov/docs/nicer/mission_guide/. Accessed: 2023-08-15.

NASA. 2013. Rossi X-ray Timing Explorer (RXTE). https://heasarc.gsfc.nasa.gov/docs/xte/XTE.html. Accessed: 2023-08-15.

NASA. 2020a. FTools HEADAS documentation: NICER Prefilter 1. https://heasarc.gsfc.nasa.gov/lheasoft/ftools/headas/niprefilter.html. Accessed: 2023-08-15.

NASA. 2020b. FTools HEADAS documentation: NICER Prefilter 2. https://heasarc.gsfc.nasa.gov/lheasoft/ftools/headas/niprefilter2.html. Accessed: 2023-08-15.

NASA. 2020c. FTools HEADAS documentation: Prefilter. https://heasarc.gsfc.nasa.gov/lheasoft/ftools/headas/prefilter.html. Accessed: 2023-08-15.

NASA. 2020d. NICER Background Estimator Tools. https://heasarc.gsfc.nasa.gov/docs/nicer/tools/nicer_bkg_est_tools.html. Accessed: 2023-08-15.

NASA. 2022. NICER data analysis threads - SCORPEON background model overview. https://heasarc.gsfc.nasa.gov/docs/nicer/analysis_threads/scorpeon-overview/. Accessed: 2023-08-15.

NASA. 2023a. HEASARC@SciServer User Guide. https://heasarc.gsfc.nasa.gov/docs/sciserver. Accessed: 2023-08-15.

NASA. 2023b. NASA's HEASoft Documentation. https://heasarc.gsfc.nasa.gov/docs/software/heasoft/. Accessed: 2023-08-15.

Okajima, T.; Soong, Y.; Balsamo, E. R.; Enoto, T.; Olsen, L.; Koenecke, R.; Lozipone, L.; Kearney, J.; Fitzsimmons, S.; Numata, A.; Kenyon, S. J.; Arzoumanian, Z.; and Gendreau, K. 2016. Performance of NICER flight x-ray concentrator. In den Herder, J.-W. A.; Takahashi, T.; and Bautz, M., eds., *Space Telescopes and Instrumentation 2016: Ultraviolet to Gamma Ray*, volume 9905, 99054X. International Society for Optics and Photonics, SPIE.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E. Z.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *CoRR*, abs/1912.01703.

Remillard, R. A.; Loewenstein, M.; Steiner, J. F.; Prigozhin, G. Y.; LaMarr, B.; Enoto, T.; Gendreau, K. C.; Arzoumanian, Z.; Markwardt, C.; Basak, A.; Stevens, A. L.; Ray, P. S.; Altamirano, D.; and Buisson, D. J. K. 2022. An Empirical Background Model for the NICER X-Ray Timing Instrument. *The Astronomical Journal*, 163(3): 130.

Tawa, N.; Hayashida, K.; Nagai, M.; Nakamoto, H.; Tsunemi, H.; Yamaguchi, H.; Ishisaki, Y.; Miller, E. D.; Mizuno, T.; Dotani, T.; Ozaki, M.; and Katayama, H. 2008. Reproducibility of Non-X-Ray Background for the X-Ray Imaging Spectrometer aboard Suzaku. *Publications of the Astronomical Society of Japan*, 60(sp1): S11–S24.

You, K.; Long, M.; Jordan, M. I.; and Wang, J. 2019. Learning Stages: Phenomenon, Root Cause, Mechanism Hypothesis, and Implications. *CoRR*, abs/1908.01878.