From Active Stars to Black Holes II: Discovery of Hα Emission in Ultra-Low Resolution *Gaia* XP Spectra with Machine Learning

Antonio C. Rodriguez[1,2], **V. Ashley Villar**[1,2], **Edo Berger**[1,2], **Valeria Villa**[3]
[1]*Center for Astrophysics | Harvard & Smithsonian, 60 Garden St. Cambridge, MA 02138*
[2]*The NSF AI Institute for Artificial Intelligence and Fundamental Interactions*
[3]*Seismological Laboratory, California Institute of Technology, Pasadena, USA.*

Abstract

In large astronomical surveys of the Milky Way, objects that feature the Hα line (6563.8 Å) in emission range from magnetically active stars to interacting binary systems which may harbor compact objects such as black holes. *Gaia* Data Release 3 (DR3) has released coefficients ("XP Spectra") encoding the full optical spectrum for 200 million sources, at least an order of magnitude greater than current spectroscopic surveys. We construct a training set of 75,789 sources with ultra-low resolution ($\lambda/\Delta\lambda \sim 50$) *Gaia* XP spectra based on ground-truth labels generated from pre-existing moderate-resolution ($\lambda/\Delta\lambda \sim 2000$) spectroscopy in the SDSS and LAMOST catalogs. Objects are either classified as showing strong Hα emission (EW < -5 Å) or not. We compare a series of classifiers based on 1) manual cuts, 2) logistic regression, 3) boosted trees (XGBoost), and deep learning with 4) multi-layer perceptrons and 5) convolutional neural networks. We find that XGBoost outperforms all other models (ROC-AUC = 0.979, F_1 = 0.810), and that all machine learning-based models outperform manual cuts (F_1 = 0.136) to detect Hα emission. We apply our best-performing XGBoost classifier to a sample of over one million unlabeled *Gaia* sources, leading to the recovery of known Hα-emitting sources and they discovery of hundreds of new interacting binary candidates. This work demonstrates the necessity of machine learning methods in discovering important sources in large, low-resolution astronomical datasets. We make the data and models publicly available for further exploration.

1 Introduction

Hα emission (6563.8 Å) is a key tracer of energetic astrophysical processes, arising in systems ranging from magnetically active stars to interacting binaries and accreting compact objects (Herbig, 1962; Warner, 1995; Reig, 2011). Large-scale identification of such sources is essential for understanding stellar activity, binary evolution, and accretion physics. Paper I (to be cited after review) presented a novel method for the identification of such sources using X-ray datasets, which are much more limited than the spectroscopic datasets for Hα detection employed here.

Traditional identification relies on moderate-resolution spectroscopy ($\lambda/\Delta\lambda \sim 2000$) from ground-based surveys such as the Sloan Digital Sky Survey (SDSS; York et al., 2000), the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST; Zhao et al., 2012), and the Dark Energy Spectroscopic Survey (DESI; DESI Collaboration et al., 2022), which provide direct measurements of Hα equivalent width. However, these surveys are limited in scale (few to ten million sources) compared to *Gaia* Data Release 3, which provides ultra low-resolution spectra ($\lambda/\Delta\lambda \sim 50$) from space for over 200 million sources (Gaia Collaboration et al., 2023; De Angeli et al., 2023). There is also a fundamental difference in the objects targeted by moderate-resolution surveys such as SDSS

and DESI compared to those in the current *Gaia* XP data release. The former surveys target faint (SDSS $g \sim 18$–20 mag) objects *outside* of the Milky Way, whereas *Gaia* is an all-sky survey with the current DR3 release of XP spectroscopy being limited to bright (*Gaia* $G < 17.6$ mag) objects mostly in the Milky Way (Gaia Collaboration et al., 2023; De Angeli et al., 2023).

Gaia XP spectra are encoded as coefficient vectors rather than sampled flux arrays, enabling efficient storage but complicating direct physical interpretation. Recent work has shown that such representations retain significant astrophysical information, particularly metallicity (Andrae et al., 2023; Leung & Bovy, 2023; Kao et al., 2024; Limberg et al., 2025; Laroche & Speagle, 2025), motivating the application of machine learning techniques to extract weak spectral signals.

In this work, we evaluate the ability of both simple physics-based cuts and machine learning models to detect Hα emission directly from Gaia XP spectra. We compare manual thresholding, linear models, neural networks, convolutional architectures, and gradient-boosted trees, and demonstrate that boosted trees achieve the best performance. We then apply our best model to a large unlabeled Gaia sample to identify new candidate emission-line systems.

2 Data and Methods

2.1 Data

We construct a labeled dataset by cross-matching *Gaia* DR3 sources that 1) have significant parallax and proper motion measurements (SNR > 5), 2) low astrometric noise (RUWE < 1.4), and 3) an XP spectrum, with spectroscopic catalogs from SDSS (York et al., 2000) and LAMOST (Zhao et al., 2012), which provide moderate-resolution spectra ($R \sim 2000$) suitable for measuring Hα equivalent widths. Sources are labeled as emission-line objects if $\mathrm{EW} < -5$ Åand non-emission-line objects if $\mathrm{EW} > 0$ Å. While this eliminates objects with weak Hα emission, the current approach ensures we have a clean sample of emission and non-emission objects. We apply all quality cuts as recommended by the SDSS and LAMOST surveys, eliminating particularly noisy spectra (SNR < 20), or were taken in poor conditions. After filtering, the dataset contains 75,789 sources with a significant class imbalance between emission (4,128) and non-emission (71,661) objects.

Gaia XP spectra are accessed via their coefficient representation, consisting of 55 BP and 55 RP coefficients per source. These are concatenated into a 110-dimensional feature vector. To remove dependence on absolute brightness, all coefficients are normalized by the Gaia `phot_g_mean_flux`. Features are then standardized to zero mean and unit variance using statistics computed on the training set only. The data are split into training, validation, and test sets using a stratified 80/10/10 split to preserve class proportions.

2.2 Models

We evaluate five classification approaches spanning interpretable baselines to modern machine learning models.

Manual Cuts. We reconstruct calibrated spectra using `GaiaXPy` (Montegriffo et al., 2023) and compute a line-to-continuum contrast at Hα:

$$\mathcal{S}_{\mathrm{H}\alpha} = \frac{F_{\mathrm{line}} - F_{\mathrm{cont}}}{F_{\mathrm{cont}}}. \tag{1}$$

The continuum is estimated from flanking wavelength windows, and a threshold (10%) is selected using the validation set, though experimentation with different thresholds revealed no improvement in ultimate classification. Ultimately, only a emission/no emission decision is made (not a probability like other methods).

Logistic Regression. We train a logistic regression classifier with ℓ_2 regularization on the normalized coefficient vectors. Class imbalance is addressed using class-weighting, assigning higher weight to the minority (emission-line) class.

Multi-Layer Perceptron (MLP). We implement a fully connected neural network with two hidden layers and ReLU activations. Dropout regularization is applied to mitigate overfitting. The model is trained using binary cross-entropy loss (`BCEWithLogitsLoss`) with a positive class weight proportional to the class imbalance.

Table 1: Performance comparison on the test set across different models.

Method	Accuracy	Precision	Recall	F_1	ROC-AUC	PR-AUC
Manual Cuts	0.945	0.203	0.102	0.136	–	–
Logistic Regression	0.888	0.460	**0.910**	0.610	0.960	0.773
MLP	0.911	0.530	0.900	0.660	0.967	0.843
1D CNN	0.906	0.510	0.900	0.650	0.969	0.858
XGBoost	**0.962**	**0.790**	0.830	**0.810**	**0.979**	**0.899**

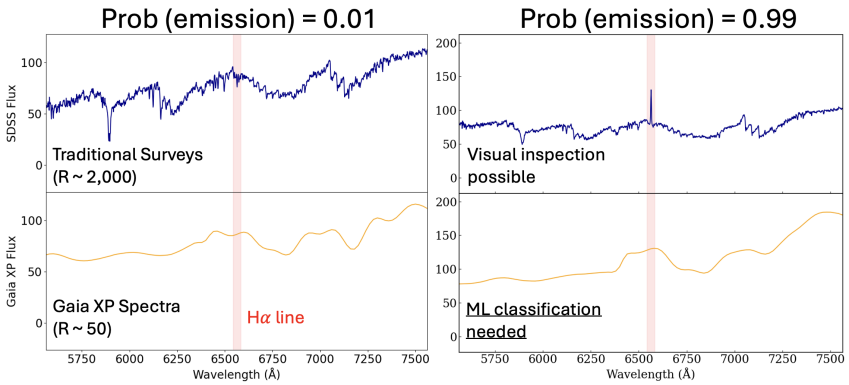

Figure 1: ML methods are a necessity in identifying Hα emission in ultra-low resolution *Gaia* XP spectra. The top panels show SDSS spectroscopy of two different M dwarf stars, one without (left) and with (right) emission. The corresponding *Gaia* XP spectra of those systems are shown in the bottom panels, that visibly resemble each other, but have different classifications generated by our XGBoost classifier.

1D Convolutional Neural Network (CNN). To explicitly exploit spectral structure, we train a 1D CNN on reconstructed flux arrays. The network consists of convolutional layers with small kernels designed to capture localized features such as emission lines, followed by pooling and fully connected layers for classification.

XGBoost. We train a gradient-boosted decision tree model (Chen & Guestrin, 2016) on the coefficient vectors. Class imbalance is handled via the `scale_pos_weight` parameter. XGBoost is particularly effective for tabular data and can capture nonlinear feature interactions without explicit feature engineering.

2.3 Evaluation

Model performance is evaluated on a held-out test set using accuracy, precision, recall, F_1 score, and both ROC-AUC and precision-recall AUC. The latter is particularly important due to class imbalance.

All models are trained on identical splits to enable direct comparison.

3 Results

3.1 Classification Performance

Table 1 shows the results for all models. All machine learning approaches significantly outperform the manual thresholding baseline, demonstrating that Gaia XP coefficient vectors encode sufficient information to identify Hα emission.

XGBoost achieves the best overall performance, with ROC-AUC = 0.979 and PR-AUC = 0.899. This indicates strong performance both in separating classes and under class imbalance. The neural network models also perform well, with the 1D CNN slightly outperforming the MLP (ROC-AUC = 0.969 vs. 0.967), suggesting that explicitly modeling spectral locality provides modest gains.

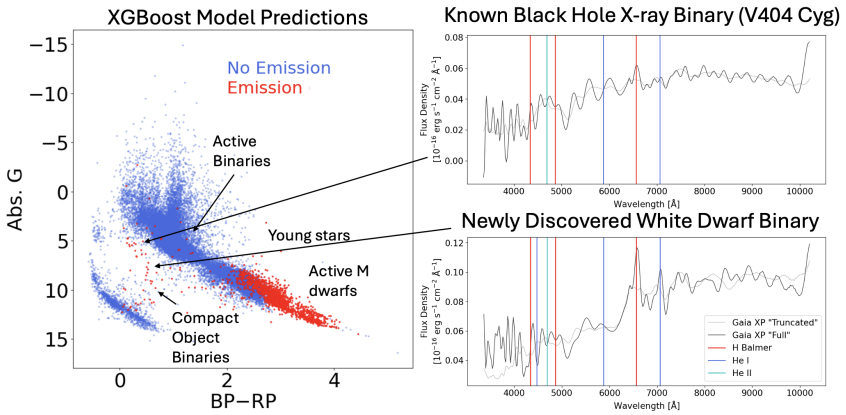

Figure 2: Applying our XGBoost classifier to a larger test set recovers known compact object binaries and reveals new candidate systems. *Left:* Model predictions on a *Gaia* Hertzsprung-Russell diagram reveal the main categories of Hα-emitting objects in expected locations. *Right: Gaia* XP spectra of known and newly discovered compact binaries reveal emission of the Hα line in addition to other lines of the Balmer series and helium.

3.2 Machine Learning Comparison to Manual Cuts

While manual cuts achieve relatively high accuracy (0.945), they suffer from extremely low recall (0.102), missing the majority of emission-line objects. This reflects the difficulty of detecting weak emission features in low-resolution spectra using simple thresholding. In contrast, machine learning models achieve both high recall (> 0.95) and substantially improved precision, indicating that they are able to leverage additional information beyond the immediate Hα region. Figure 1 shows the difficulty in visually identifying Hα emission at the ultra-low resolution of *Gaia*, while our XGBoost classifier correctly identifies Hα emission with high confidence.

The strong performance of coefficient-based models suggests that information relevant to Hα emission is distributed across the spectrum. Models likely exploit correlations with continuum shape, stellar temperature, and other spectral features associated with chromospheric activity and accretion processes (West et al., 2011; Reiners et al., 2012). Furthermore, Figure 1 also suggests that at the low resolution of *Gaia* XP spectra, Hα emission may only be easily detectable with machine learning methods, at least in some sources.

4 Conclusion

We demonstrate that Hα emission can be reliably detected in Gaia XP spectra using machine learning, despite their ultra-low resolution. All ML models outperform manual thresholding, with XGBoost achieving the best performance (ROC-AUC = 0.979).

These results show that Gaia XP coefficients encode rich astrophysical information that can be extracted without explicit spectral reconstruction. Applying our classifier to unlabeled Gaia data enables efficient discovery of new emission-line systems, including candidate interacting binaries. We speculate that XGBoost achieved a superior performance to all models due to its more fine-grained approach to classification, and possibly its ability to build decision trees from non-neighboring spectral coefficients. Further interpretability analysis will be carried out to determine what spectral coefficients had the highest impact on the ultimate classification.

Future work will apply these models to the full set of 200 million *Gaia* XP spectra, including the release of billions of spectra in Data Release 4 and beyond. This work will be extended to cover additional emission lines, including unsupervised approaches which may be able to identify emission without a parent labeled dataset from SDSS or LAMOST. In sum, this work has revealed the necessity of machine learning usage in large astronomical surveys to extract information from low-resolution data that may have gone ignored otherwise. We make the data and models publicly available at `https://github.com/acrodrig98/EmissionGaiaXP`.

References

Gaia Collaboration, Prusti, T., et al. 2016, *A&A*, 595, A1

Gaia Collaboration, Vallenari, A., et al. 2023, *A&A*, 674, A1

De Angeli, F., et al. 2023, *A&A*, 674, A2

Montegriffo, P., et al. 2023, *A&A*, 674, A3

Andrae, R., et al. 2023, *A&A*, 674, A27

York, D. G., et al. 2000, *AJ*, 120, 1579

Zhao, G., et al. 2012, *RAA*, 12, 723

DESI Collaboration, et al. 2022, *AJ*, 164, 207

Herbig, G. H. 1962, *ApJ*, 135, 736

Joy, A. H. 1945, *ApJ*, 102, 168

Warner, B. 1995, *Cataclysmic Variable Stars* (Cambridge: Cambridge Univ. Press)

Reig, P. 2011, *Ap&SS*, 332, 1

West, A. A., et al. 2011, *AJ*, 141, 97

Frasca, A., et al. 2015, *A&A*, 575, A4

Reiners, A., et al. 2012, *ApJ*, 746, 43

Ness, M., et al. 2015, *ApJ*, 808, 16

Ting, Y.-S., et al. 2019, *ApJ*, 879, 69

Guiglion, G., et al. 2020, *A&A*, 640, A111

Leung, H. W., & Bovy, J. 2019, *MNRAS*, 483, 3255

Leung, H. W., & Bovy, J. 2023, *MNRAS*

Chen, T., & Guestrin, C. 2016, Proceedings of the 22nd ACM SIGKDD Conference

Laroche, A. & Speagle, J. S. 2025, ApJ, 979, 1, 5. doi:10.3847/1538-4357/ad9607

Limberg, G., Placco, V. M., Ji, A. P., et al. 2025, ApJL, 989, 1, L18. doi:10.3847/2041-8213/adf196

Kao, M. L., Hawkins, K., Rogers, L. K., et al. 2024, ApJ, 970, 2, 181. doi:10.3847/1538-4357/ad5d6e