

Hybrid Machine Learning Classification of AGB Stars in Gaia DR3

Tanya Kushwahaa^{id}, Mikako Matsuura^{id}, Roger Wesson and Timothy A. Davis

Cardiff Hub for Astrophysics Research and Technology (CHART), School of Physics and Astronomy, Cardiff University, The Parade, Cardiff CF24 3AA, UK
email: tkushwahaa99@gmail.com

Abstract. Machine learning (ML) is revolutionising the analysis of large astronomical surveys such as *Gaia* DR3. We present a two-step ML pipeline to classify Asymptotic Giant Branch (AGB) stars: (i) cosine similarity identifies candidate AGB stars from 140 million sources, and (ii) an XGBoost classifier which separates carbon-rich (C-rich) and oxygen-rich (O-rich) stars. Our classifier achieves ROC-AUC > 99.9%, producing a catalogue of 307,191 O-rich and 4,227 C-rich stars. Feature importance analysis shows that the sixth RP coefficient (RP6) dominates classification, underlining the critical role of *Gaia* red spectral features. This work demonstrates how hybrid ML approaches can efficiently identify chemically distinct stellar populations in massive survey datasets, providing opportunities for new studies of Galactic structure and stellar evolution.

Keywords. Machine Learning, Stellar Classification, AGB Stars, Gaia

1. Introduction

Understanding the formation and evolution of the Milky Way remains a fundamental challenge in Galactic astronomy. The *Gaia* mission ([Gaia Collaboration 2016](#)) has revolutionised this field, with Data Release 3 (DR3) providing low-resolution BP/RP spectra for 220 million stars ([Gaia Collaboration 2023](#)).

Asymptotic Giant Branch (AGB) stars, tracing intermediate-age populations (0.1–10 Gyr), are key probes of Galactic evolution. They contribute significantly to dust production and chemical enrichment, and their carbon-rich (C-rich) and oxygen-rich (O-rich) subclasses provide insights into stellar evolution and metallicity gradients ([Vassiliadis & Wood 1993](#); [Marigo et al. 2017](#)). However, identifying AGB stars in large photometric and spectroscopic catalogues is challenging due to overlap with other stellar populations and the limited availability of labelled data.

Machine learning (ML) offers a powerful solution for high-dimensional classification tasks in astronomy ([Sanders & Matsunaga 2023](#); [Liu & Xu 2024](#)). In this work, we develop a hybrid two-step ML pipeline: first, similarity analysis selects candidate AGB stars by comparison with known catalogues; second, a supervised XGBoost classifier distinguishes C-rich and O-rich populations using combined photometric and *Gaia* XP spectral coefficients. This method effectively addresses class imbalance and scales to millions of sources, allowing us to construct a comprehensive Galactic AGB catalogue.

2. Methodology

2.1. Data and Features

We constructed our dataset by cross-matching *Gaia* DR3 sources with XP spectra against the Two Micron All Sky Survey (2MASS) (Skrutskie *et al.* 2006) and the Wide-field Infrared Survey Explorer (WISE) (Wright *et al.* 2010), creating the *G2MW* sample of 140 million stars. For training, we compiled known AGB stars from the General Catalogue of Galactic Carbon Stars (CGCS) and the General Catalogue of Variable Stars (GCVS), resulting in 1,328 C-rich and 3,680 O-rich stars after quality filtering.

Our feature set combines nine colour indices (*Gaia*, 2MASS, WISE) with 110 BP/RP spectral coefficients. These features capture the key optical-to-infrared signatures of AGB stars: C-rich stars display strong C₂ and CN molecular bands, while O-rich stars show prominent TiO absorption features.

2.2. Two-step Classification Pipeline

Traditional supervised and unsupervised methods proved inadequate for the heterogeneous *Gaia* dataset. We developed a hybrid approach combining similarity analysis with supervised classification:

Step 1: Similarity Analysis - We first applied cosine similarity to identify candidate AGB stars from the *G2MW* sample by comparing their spectral and colour features to known AGB stars. This metric evaluates alignment in high-dimensional feature space between each *Gaia* star and known AGB stars:

$$S_c = \cos \theta = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|} \quad (1)$$

Stars with the highest similarity scores form a refined candidate set which are passed to the classifier.

Step 2: XGBoost Classification We trained a supervised XGBoost classifier on the known AGB sample (~5,000 sources) to distinguish C-rich and O-rich stars. Class imbalance (O-rich:C-rich ~ 2.8:1) is handled using the `scale_pos_weight` parameter in XGBOOST. We use a stratified 70-30 train-test split, five-fold cross-validation, and hyperparameter tuning via `RandomizedSearchCV`. We use ROC-AUC (Receiver Operating Characteristic Area Under the Curve) as the evaluation metric, which measures classification performance across all probability thresholds, and L1/L2 regularisation techniques to prevent overfitting.

This two-step approach (Fig. 1) ensures accurate classification by training on well-labelled data while applying to a carefully selected candidate sample.

3. Results

Our similarity analysis identified approximately 533,000 AGB candidates from the 140 million *G2MW* sources using a 99.999th percentile threshold. This threshold balances completeness (capturing diverse AGB evolutionary phases) with purity (minimising contamination from non-AGB sources).

The XGBOOST classifier achieved excellent performance across different feature combinations. Using spectroscopic features alone yielded ROC-AUC > 99.9% and accuracy > 99.6% on the test set. Colour-only features achieved 98% accuracy, while combined features provided the most robust classification with 0.3% error rate.

Feature importance analysis highlights that the sixth RP coefficient dominates classification, contributing over 40% of the combined model's predictive power (Fig. 2), followed by $J - K$ colour (5-6%). This confirms that red spectral features corresponding

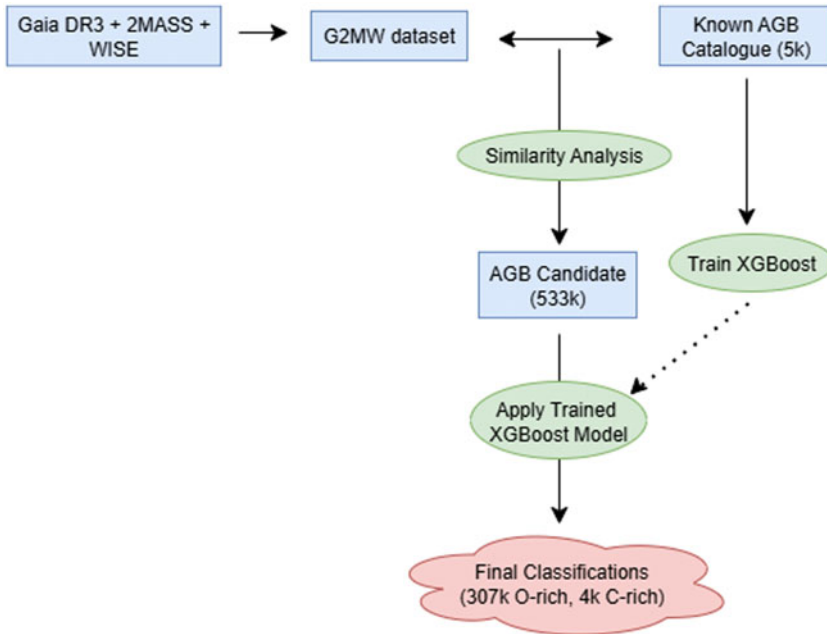


Figure 1. Schematic representation of our two-step machine learning pipeline for classifying C-rich and O-rich AGB stars from *Gaia* DR3. The pipeline combines data from multiple surveys (*Gaia* DR3 + 2MASS + WISE → *G2MW* dataset) with known AGB catalogues (General Catalog of Galactic Carbon Stars (CGCS) and General Catalog of Variable Stars (GCVS)). Step 1 uses cosine similarity analysis to identify 533k AGB candidates from the *G2MW* dataset by comparison with known AGB stars. Step 2 trains an XGBOOST classifier on the known AGB sample (5k sources) and applies the trained model to classify the AGB candidates, yielding 307k O-rich and 4k C-rich AGB stars.

to molecular absorption bands spanning 630–1100nm are crucial for distinguishing C-rich and O-rich chemistry in AGB stars.

Applying probability threshold $p > 0.95$, we classified 307,191 O-rich and 4,227 C-rich AGB stars from *Gaia* DR3. This represents a substantial expansion of the known Galactic AGB population.

4. Validation and Astrophysical Insights

Visual inspection of *Gaia* XP spectra confirms our classifications (see Fig. 3b, c). C-rich AGB stars display multiple C₂ and CN molecular absorption bands, while O-rich AGB stars show strong TiO and VO features. The absolute M_{K_s} vs. $J - K_s$ colour-magnitude diagram (Fig. 3a) further validates our results, with C-rich stars peaking at $J - K_s \sim 1.63$ and O-rich stars at $J - K_s \sim 1.09$, consistent with established AGB classification schemes (Cioni & Habing 2003).

Wesenheit function analysis validates our classification (Lebzelter et al. 2018), with 62% of classified C-rich stars occupying the expected C-rich region and 52% of O-rich stars in the low-mass O-rich branch. We identify O-rich stars with circumstellar dust that appear in C-rich regions of the Wesenheit diagram; however, visual inspection of their spectra confirms they are O-rich AGB stars, highlighting the advantages of spectroscopic validation over purely photometric methods.

Spatial distribution analysis shows distinct population patterns: O-rich AGB stars concentrate near the Galactic plane and center ($|b| \pm 15^\circ$), tracing the older stellar populations. In contrast, C-rich stars have a more extended distribution ($|b| \pm 10^\circ$), spanning

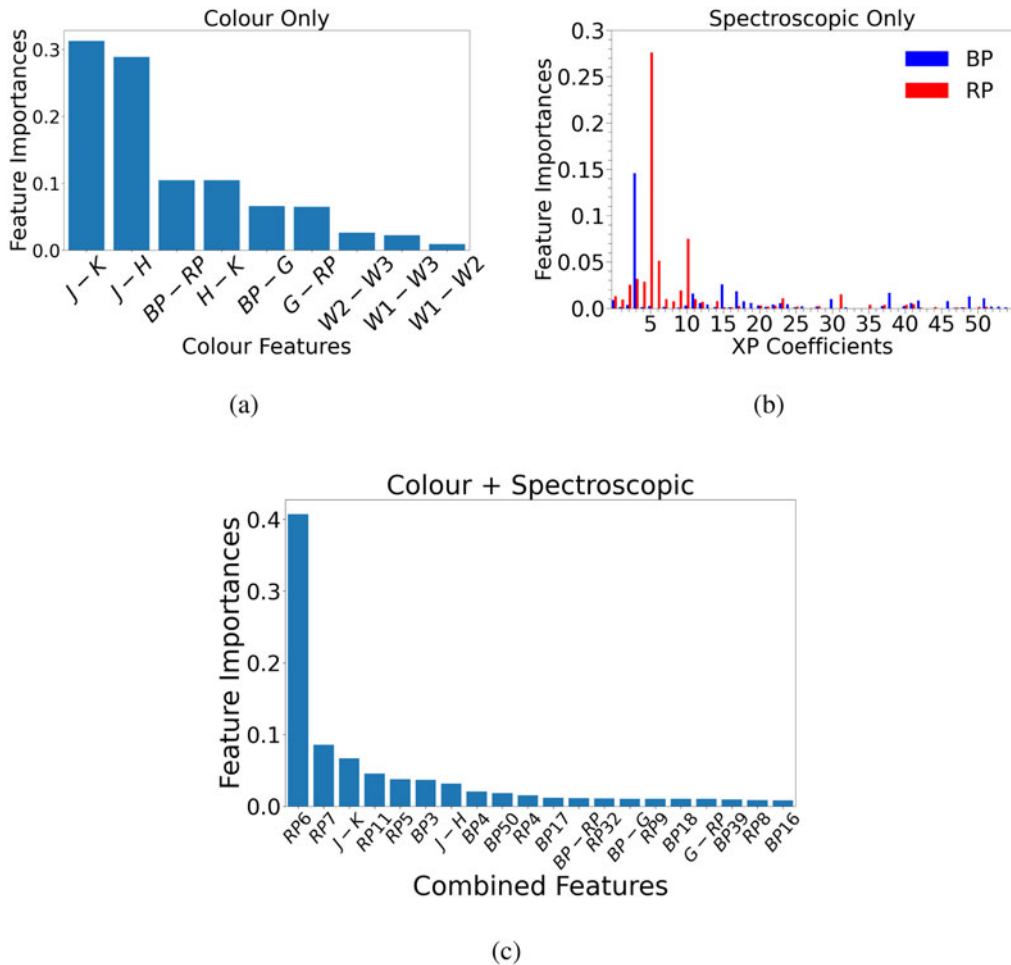


Figure 2. XGBoost feature importance for (a) colour, (b) spectroscopic, and (c) combined features. The sixth RP coefficient (RP6) dominates classification (contributing >40% in the combined model), followed by $J - K$ colour, demonstrating the value of red spectral information for AGB chemistry.

the full Galactic longitude range, with concentrations in the spiral arms and the Galactic anticenter. We identify 409 C-rich stars within $|l| \pm 30^\circ$ of the Galactic center, nearly doubling previous bulge surveys.

5. Conclusion

We present a hybrid two-step ML pipeline combining cosine similarity and XGBOOST to classify AGB stars from *Gaia* DR3. The method achieves ROC-AUC > 99.9% and produces a comprehensive catalogue of 307,191 O-rich and 4,227 C-rich stars from *Gaia* DR3.

Key contributions include effective candidate selection from a dataset of 140 million stars, robust handling of class imbalance in AGB populations, and demonstrating that red spectral coefficients, particularly RP6, are critical for distinguishing C-rich and O-rich AGB stars.

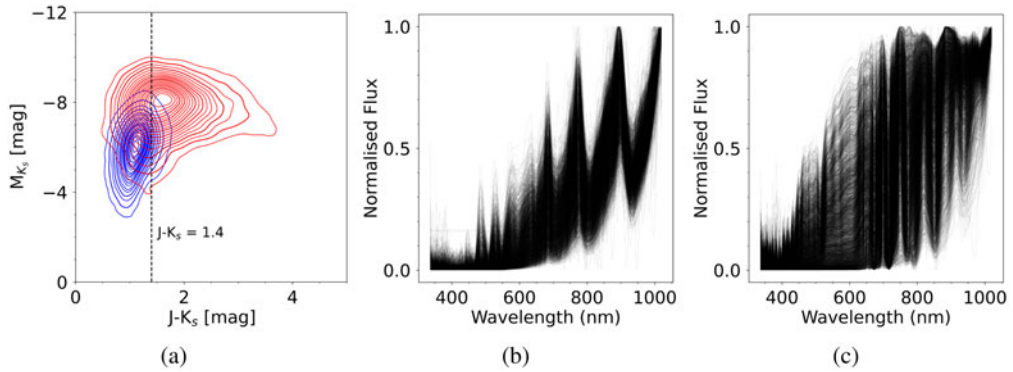


Figure 3. Classification validation results. (a) Absolute K_s magnitude vs $J - K_s$ colour-magnitude diagram showing ML-classified C-rich (red contours) and O-rich (blue contours) AGB stars. Contour levels represent stellar density distributions. The dashed vertical line at $J - K_s = 1.4$ mag shows the traditional photometric separation boundary (Cioni & Habing 2003). (b) Representative *Gaia* XP spectra of classified C-rich AGB stars showing characteristic C_2 and CN molecular absorption bands. (c) Representative *Gaia* XP spectra of classified O-rich AGB stars displaying prominent TiO and VO molecular absorption bands.

This work demonstrates how machine learning can efficiently process large astronomical datasets, opening new avenues for studying stellar evolution, chemical enrichment, and Galactic structure with current and future survey data.

Data Availability Statement

The sample selected by the machine-learning method will be made available alongside the main paper (Kushwahaa et al. 2025, submitted).

References

- Cioni, M.-R. L., & Habing, H. J. 2003, *A&A*, 402, 133
- Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., *et al.* 2016, *A&A*, 595, A1
- Gaia Collaboration, Vallenari, A., Brown, A. G. A., *et al.* 2023, *A&A*, 674, A1
- Lebzelter, T., Mowlavi, N., Marigo, P., *et al.* 2018, *A&A*, 616, L13
- Liu, A., & Xu, J. C. 2024, *RNAAS*, 8, 83
- Marigo, P., Girardi, L., Bressan, A., *et al.* 2017, *ApJ*, 835, 77
- Sanders, J. L., & Matsunaga, N. 2023, *MNRAS*, 521, 2745
- Skrutskie, M. F., Cutri, R. M., Stiening, R., *et al.* 2006, *AJ*, 131, 1163
- Vassiliadis, E., & Wood, P. R. 1993, *ApJ*, 413, 641
- Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., *et al.* 2010, *AJ*, 140, 1868