أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

The Iron in Their Rhythm

Machine-Learned Metallicities for RR Lyrae Stars from Photometric Light Curves

SUDS Research Internship — Final Report

Yazan Bakhshwin

KAUST Academy · King Abdulaziz University

Supervisors: Prof. Ting Li and Dr. Gustavo Medina

Internship host: SUDS Program, Data Sciences Institute, University of Toronto

Sponsored by KAUST Academy

yazan.bakhshwin@kaust.edu.sa

Abstract

The metallicity ([Fe/H]) of stars in the Milky Way provides a fossil record of its assembly, and the ancient RR Lyrae variables are among its best tracers. Measuring metallicity traditionally requires spectroscopy, which is expensive and available for only a small fraction of stars, while photometric surveys deliver light curves for millions. This report summarizes an internship project that bridges the two: models were trained to predict spectroscopic [Fe/H] from the shape of the photometric light curve alone, using RR Lyrae with DESI spectroscopic labels cross-matched to ZTF photometry, with the RRab and RRc subtypes modeled separately. Gradient-boosted trees on engineered light-curve shape features proved the strongest of the model families compared, roughly halving the error of a guess-the-mean baseline; a data-quality investigation identified unreliable pulsation periods as the dominant error source and addressed them with a wider period search and a reliability classifier. The trained models were then applied to every Gaia DR3 RR Lyrae with complete published light-curve parameters, producing a catalog of photometric metallicities for over 135,000 stars, roughly thirty times the spectroscopic training sample. Validated against globular clusters and dwarf galaxies whose metallicities never entered training, the catalog agrees with the literature at approximately the 0.15 dex level and compares favorably with the Gaia mission's own photometric metallicities on the same systems. The report also documents two instructive negative results: the scarcity of metal-rich training labels sets an information limit that modeling choices do not overcome, and self-supervised pretraining on the full unlabeled corpus adds no information beyond the engineered features. Together these results deliver a validated, uncertainty-quantified metallicity catalog built from photometry alone, with a clear account of where it can be trusted.

Date: August 2026

Contents

1 Introduction and Scientific Motivation **2**

2 Related Work **2**

3 Data **3**

3.1 ZTF Light Curves . . . 3

3.2 DESI Spectroscopic Labels . . . 3

3.3 Quality Cuts and Sample Sizes . . . 4

4 Methods **4**

4.1 Feature Extraction . . . 4

4.2 Models Compared . . . 5

4.3 Evaluation Protocol . . . 5

5 Results: Model Comparison **5**

5.1 The Model Ladder . . . 5

5.2 Headline Accuracy . . . 5

5.3 The Metal-Rich Regime . . . 6

6 Data-Quality Investigation: Periods **7**

6.1 Period Disagreements and the Wrong-Period Stress Test . . . 7

6.2 Recovering Aliased Short-Period RRc . . . 8

6.3 A Period-Reliability Classifier . . . 9

7 The Catalog: 135,873 Stars, All-Sky **9**

8 External Validation **10**

8.1 Globular Clusters and Dwarf Galaxies . . . 10

8.2 Auditing the Per-Star Uncertainties . . . 10

9 Self-Supervised Pretraining: A Measured Negative Result **12**

10 Limitations **13**

11 Conclusions and Future Work **13**

Acknowledgments **14**

References **14**

1 Introduction and Scientific Motivation

The chemical abundances of individual stars encode the formation history of the Galaxy, and metallicity [Fe/H] (the logarithmic iron-to-hydrogen ratio relative to the Sun, measured in dex) is the single most informative such quantity for tracing the assembly of the Milky Way halo. RR Lyrae stars (old, luminous, radially pulsating horizontal-branch variables) are exceptional tracers: they are standard candles, they are numerous in the halo, and their pulsation properties are known to depend on their iron abundance. A metallicity for each RR Lyrae therefore tags every point of a halo map with chemical information, and structures such as accreted streams and disrupted satellites become separable by their abundances.

The obstacle is that metallicity is normally read from spectra, which are observationally expensive and exist for only a small fraction of known RR Lyrae. Time-domain photometric surveys, by contrast, record light curves (repeated brightness measurements) for these stars by the million. Because the iron content changes how an RR Lyrae pulsates, [Fe/H] is imprinted in the period and shape of the light curve, and the problem can be posed as supervised regression: given a light curve, predict the spectroscopic metallicity. This project asks that question with a practical payoff in mind: can a few thousand spectroscopic labels train a model that delivers trustworthy photometric metallicities for the roughly 10^5 known RR Lyrae, and can the resulting predictions be validated star by star?

This final report consolidates the full arc of the internship, from raw survey data to a delivered and externally validated data product. Its main contributions are:

1. a labeled dataset of RR Lyrae with DESI spectroscopic metallicities cross-matched to ZTF light curves, with quality cuts and the RRab and RRc subtypes treated separately throughout;
2. a controlled comparison of model families (gradient-boosted trees, random forests, multilayer perceptrons, and convolutional networks) under a repeated-split evaluation protocol with no tuning on the test set, in which XGBoost on engineered light-curve shape features emerges as the strongest model at this label count;
3. a data-quality investigation showing that unreliable pulsation periods are the dominant error source, quantified with a deliberate wrong-period stress test and mitigated with a wider period search and a period-reliability classifier;
4. an all-sky application of the pipeline, yielding a catalog of 135,873 photometric metallicities, roughly 33× the spectroscopic training sample;
5. external validation against globular clusters and dwarf galaxies whose metallicities never entered training, including an audit of the per-star uncertainties on the same ground truth; and
6. two documented negative results, on mitigating the metal-rich label scarcity and on self-supervised pretraining at scale, which we consider part of the scientific value of the project.

The remainder of the report follows this structure: data (Section 3), methods (Section 4), model results (Section 5), the period-quality investigation (Section 6), the production catalog (Section 7), external validation (Section 8), self-supervised pretraining (Section 9), and limitations and conclusions (Sections 10–11).

2 Related Work

Photometric metallicities for RR Lyrae have a long history rooted in the empirical relations between [Fe/H] and Fourier parameters of the light curve [1], and more recent work has applied

machine learning to large survey samples. Dékány et al. [2] derived photometric metallicities from I-band Fourier decompositions with machine-learned models, providing a natural external benchmark for the transfer of our relations to an independent survey. Gaia DR3 publishes its own photometric metallicities for RR Lyrae, computed from Gaia light-curve Fourier parameters [3]; these serve as a baseline for the same stars but must be treated carefully, since a model trained on the same Gaia features reproduces that column almost perfectly by construction rather than by genuine skill.

For the external validation presented here we rely on the homogeneous globular cluster metallicity scale of Harris [4] and on cluster membership built from Gaia astrometry [5]. Validating photometric metallicities against globular clusters is standard practice in the field: each cluster contributes many member stars at essentially a single, carefully measured abundance, providing ground truth that is independent of the spectroscopic survey used for training.

Our work differs from previous large-scale efforts in combining a validated, survey-standalone pipeline with per-star uncertainties that are themselves audited against the cluster ground truth, and in reporting systematic negative results (on bias mitigation and on self-supervised pretraining) alongside the catalog.

3 Data

3.1 ZTF Light Curves

The photometric inputs are light curves from the Zwicky Transient Facility (ZTF) [6], a wide-field survey that has imaged the northern sky repeatedly since 2018. A typical star in our sample has hundreds to a few thousand brightness measurements (epochs) in the g and r bands spread over about seven years. The raw time series looks like scatter (Figure 1, left): the pulsation period of an RR Lyrae (typically 0.2–0.9 days) is far shorter than the gaps between observations. Folding the measurements by the pulsation period aligns them into the characteristic curve shape that carries the physical information. The two RR Lyrae subtypes are modeled separately throughout this work because their shapes differ qualitatively: RRab stars (fundamental-mode pulsators, longer periods) show a sharp-rise, slow-decline sawtooth, while RRc stars (first-overtone pulsators, shorter periods) show a smoother, nearly sinusoidal curve (Figure 1, center and right).

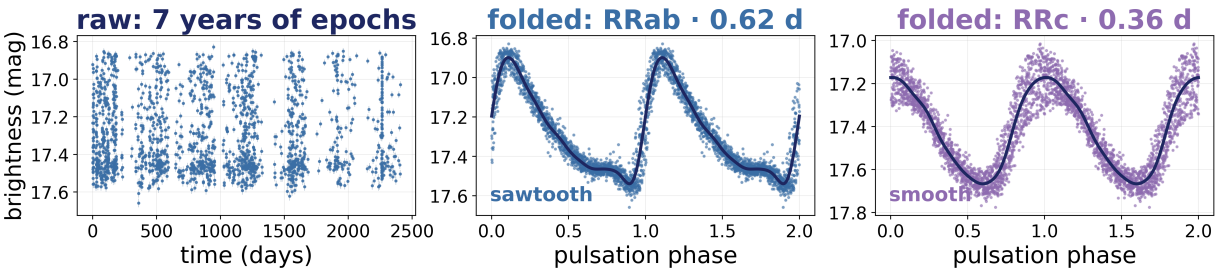


Figure 1 From raw photometry to model input. Left: seven years of ZTF epochs for one star; the pulsation is invisible at this scale. Center: the same star folded at its 0.62-day period, revealing the sawtooth shape of an RRab. Right: a folded RRc, smoother and shorter-period. The models read features of these folded shapes.

3.2 DESI Spectroscopic Labels

The training labels are spectroscopic metallicities from the Dark Energy Spectroscopic Instrument (DESI) [7], whose early data include [Fe/H] measurements (with per-star uncertainties) for several

thousand RR Lyrae. Cross-matching the DESI RR Lyrae against ZTF photometry yields a labeled sample of roughly 4,000 stars, which is the entire supervised signal available to this project: every one of the $\sim 10^5$ catalog predictions in Section 7 traces back to these labels.

3.3 Quality Cuts and Sample Sizes

The labeled sample is cleaned before modeling: stars must have a valid variability classification (classifier score above 0.5), non-missing [Fe/H], label uncertainty, period, and amplitude, and a light curve sufficiently sampled for stable feature extraction. Cuts on the spectroscopic label uncertainty were explored at several thresholds (FEH_ERR < 0.3, 0.2, 0.15); the headline results are not sensitive to this choice, and the label uncertainties themselves are treated with an error floor when used for weighting. After the period-reliability screening described in Section 6, the modeling sample contains 2,859 RRab and 996 RRc stars.

A property of the labels that shapes much of what follows is their distribution (Figure 2): the sample is dominated by metal-poor halo stars, and only 9% of the training labels (355 stars) lie on the metal-rich side of [Fe/H] = −1.3. The consequences of this imbalance are examined in Section 5.3.

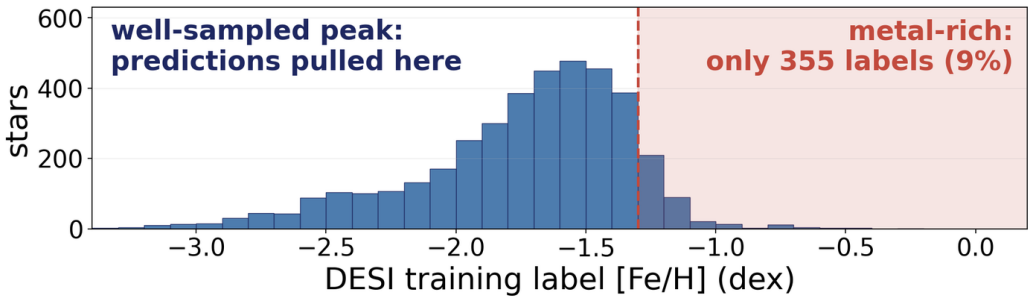


Figure 2 The DESI training-label distribution. The metal-rich side of the dashed line contains 355 stars, 9% of the sample: the model has almost nothing to learn from in the regime where it turns out to be weakest.

4 Methods

4.1 Feature Extraction

The features are engineered descriptions of the folded light-curve shape, computed with the ZTF feature-extraction pipeline developed by our collaborators Nina Hernitschek and Sergey Koposov, which we verified end-to-end against an independent re-implementation. For each star the pipeline performs a periodogram search for the pulsation period and then fits a truncated Fourier (harmonic) series to the folded curve,

$$m(\phi) \;=\; m_0 \;+\; \sum_{k=1}^{K} A_k \cos\!\big(2\pi k\phi + \varphi_k\big)\,, \tag{1}$$

so that the shape is summarized by the harmonic amplitudes and phases. The model inputs are the pulsation period, the g-band peak-to-peak amplitude, and the low-order harmonic coefficients, including the amplitude ratios and phase differences (R_{21}, R_{31}, φ_{21}, φ_{31}, and related terms) that classical work has long connected to metallicity [1]. We use the harmonic-coefficient parameterization throughout, having found it numerically more stable than an alternative free Fourier optimization in the same pipeline.

4.2 Models Compared

Five model families were evaluated on identical features and splits, always separately for RRab and RRc:

- **Guess-the-mean null**: predicts the training-set mean [Fe/H] for every star; the floor any useful model must beat.
- **XGBoost on period + amplitude only**: a two-feature baseline that isolates how much of the signal the full curve shape adds.
- **Random forest** and **XGBoost** (gradient-boosted trees) on the full engineered shape features.
- **Multilayer perceptron (MLP)** on the same shape features.
- **Convolutional neural network (CNN)** reading the folded light curve directly as an image, with no engineered features.

4.3 Evaluation Protocol

All results follow one protocol, fixed early in the project and never relaxed:

- an 80/20 train–test split at a fixed random seed, with hyperparameters tuned by cross-validation on the training split only; the test set is never used for tuning or model selection;
- headline numbers averaged over 15 random splits, so that no result depends on one lucky partition; quoted uncertainties are the standard deviation across splits;
- metrics: mean absolute error (MAE, in dex), the coefficient of determination R^2, and bias (mean of predicted minus true), each reported per subtype.

For interpretation it helps to note that one dex is a factor of ten in iron abundance, and that the spectroscopic labels themselves carry measurement noise, so the measured model error includes the labels' own noise and is in that sense an upper bound on the true model error.

5 Results: Model Comparison

5.1 The Model Ladder

Table 1 and Figure 3 summarize the comparison of model families under the repeated-split protocol. The pattern is consistent across both subtypes. Guessing the mean sets the floor at 0.29 dex (RRab) and 0.37 dex (RRc). Period and amplitude alone recover only a modest part of the signal; the large step comes from the harmonic shape features, confirming that the chemical information lives in the *shape* of the folded curve. Among models reading those features, gradient-boosted trees are the most accurate and the most stable: XGBoost reaches 0.16 dex (RRab) and 0.19 dex (RRc), with random forests close behind. The CNN, reading the folded curve as an image with no engineered features, comes genuinely close, which we take as independent evidence that the shape carries the signal; but it does not surpass the feature-based trees at this label count. The MLP on the same features is markedly less stable, especially for the smaller RRc sample, where individual splits diverge badly.

5.2 Headline Accuracy

Figure 4 shows predicted versus spectroscopic [Fe/H] for the adopted XGBoost models, with every labeled star plotted at its held-out prediction pooled across all 15 splits, so the figure contains no lucky-split selection. The 15-split averages are MAE = 0.16 dex, $R^2 = 0.65$ for RRab and MAE = 0.19 dex, $R^2 = 0.68$ for RRc. The overall bias is consistent with zero for both subtypes ($+0.00 \pm 0.01$ dex RRab, -0.00 ± 0.02 dex RRc), although Section 5.3 shows that this global balance

Table 1 Held-out MAE (dex) by model family, mean ± standard deviation over 15 random 80/20 splits. Sample sizes: 2,859 RRab (≈ 2,287 train / 572 test per split) and 996 RRc (≈ 797 train / 199 test per split). Best per column in bold.

Model	RRab	RRc
Guess the mean (null)	0.29 ± 0.01	0.37 ± 0.02
XGBoost: period + amplitude only	0.28 ± 0.01	0.29 ± 0.01
MLP: shape features	0.21 ± 0.02	0.63 ± 0.33
CNN: folded-curve image	0.17 ± 0.01	0.21 ± 0.01
Random forest: shape features	0.17 ± 0.01	0.19 ± 0.01
XGBoost: shape features	**0.16 ± 0.01**	**0.19 ± 0.01**

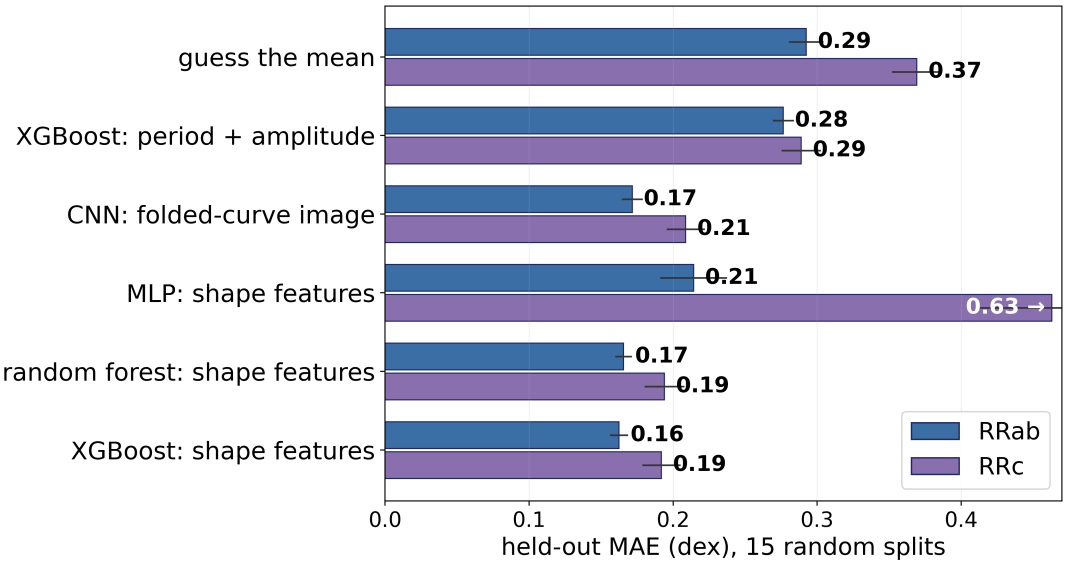


Figure 3 The model ladder: held-out MAE over 15 random splits, RRab in blue and RRc in purple (lower is better). The off-scale RRc MLP bar is clipped with its value labeled. Curve-shape features with gradient-boosted trees win at this label count.

hides a metallicity-dependent structure. In practical terms, a typical prediction lands within about 0.16 dex of what the spectrograph would measure, roughly half the guess-the-mean error; and since the labels carry their own measurement noise, this figure is an upper bound on the model error. Training takes minutes on a laptop, and labeling the full catalog of Section 7 takes seconds.

5.3 The Metal-Rich Regime

The predictions are visibly compressed toward the sample mean at the metal-rich end: stars above [Fe/H] ≈ −1.3 are systematically predicted too metal-poor, with the bin-level bias reaching about −0.3 dex. Given the label distribution of Figure 2 (9% of labels in this regime), we ran a systematic mitigation program under the same repeated-split protocol: (i) one-sided reweighting of metal-rich training stars, (ii) asymmetric loss functions penalizing under-prediction of metal-rich stars, (iii) post-hoc decompression of the predicted distribution, and (iv) bin-wise residual corrections. One-sided reweighting gives the best trade-off, improving the metal-rich test bin from 0.30 to 0.22 dex at a cost of less than 0.01 dex overall; post-hoc decompression can flatten the

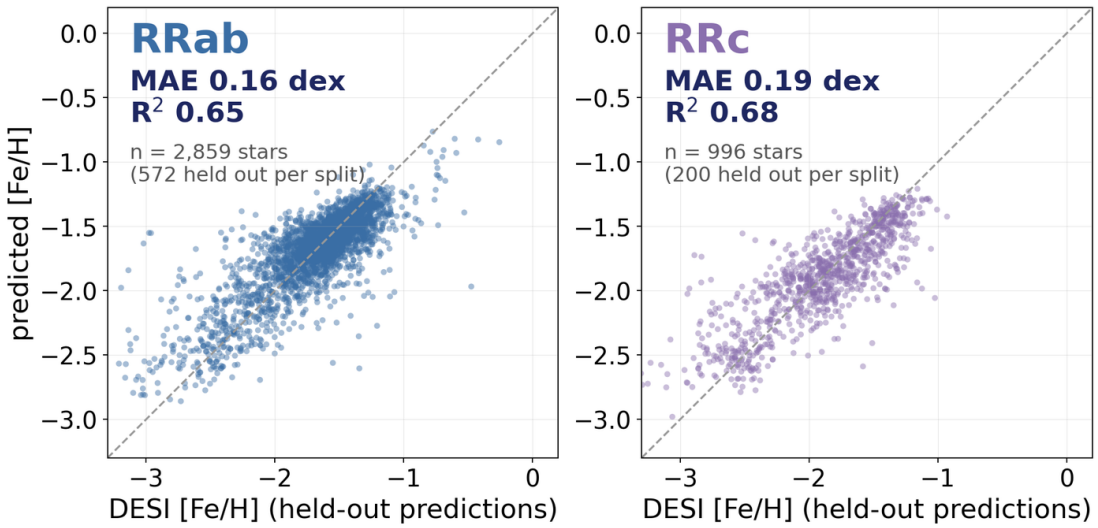


Figure 4 Predicted versus spectroscopic [Fe/H] on held-out data. Every star in the labeled sample appears at the prediction it received when it was in the test set, averaged over the ~3 of 15 splits that held it out; the quoted MAE and R^2 are means over the per-split scores.

metal-rich bias almost entirely (from −0.29 to −0.02 dex) but costs 0.06 dex in overall MAE and overshoots on metal-poor stars. Crucially, the compression slope itself barely moves under any arm: the different methods trade bias for variance without adding information. We conclude, with reasonable confidence, that the limitation is the label distribution rather than the modeling, and that its resolution is new metal-rich spectroscopy; the affected stars are flagged in the released catalog.

6 Data-Quality Investigation: Periods

Every feature the models read is computed from the folded curve, so the entire pipeline stands on the assumed pulsation period. This section summarizes the investigation that made period quality the project's central data-quality theme.

6.1 Period Disagreements and the Wrong-Period Stress Test

Comparing the ZTF-derived periods against the independently derived Gaia periods for the same stars revealed that a substantial minority (roughly a fifth of the sample) disagree beyond tolerance: irregular sampling can make a plausible but wrong period fit the data (aliasing), and blended nearby stars can distort the curve. To measure what a wrong period costs, we ran a deliberate stress test: the same trained models were fed features computed at a knowingly wrong (aliased) period and compared with the same stars processed at a consistent period (Figure 5). The effect is dramatic: MAE inflates by a factor of roughly three to five, from 0.16 to 0.42 dex (XGBoost, RRab) and 0.16 to 0.63 dex (XGBoost, RRc), with the CNN degrading even more steeply (0.16 to 0.72 and 0.21 to 1.20 dex respectively). In every case the wrong-period error exceeds the guess-the-mean baseline on the same stars: an undetected wrong period is worse than having no light curve at all. The same stars processed at a consistent period behave normally, which localizes the damage to the period itself rather than to the stars.

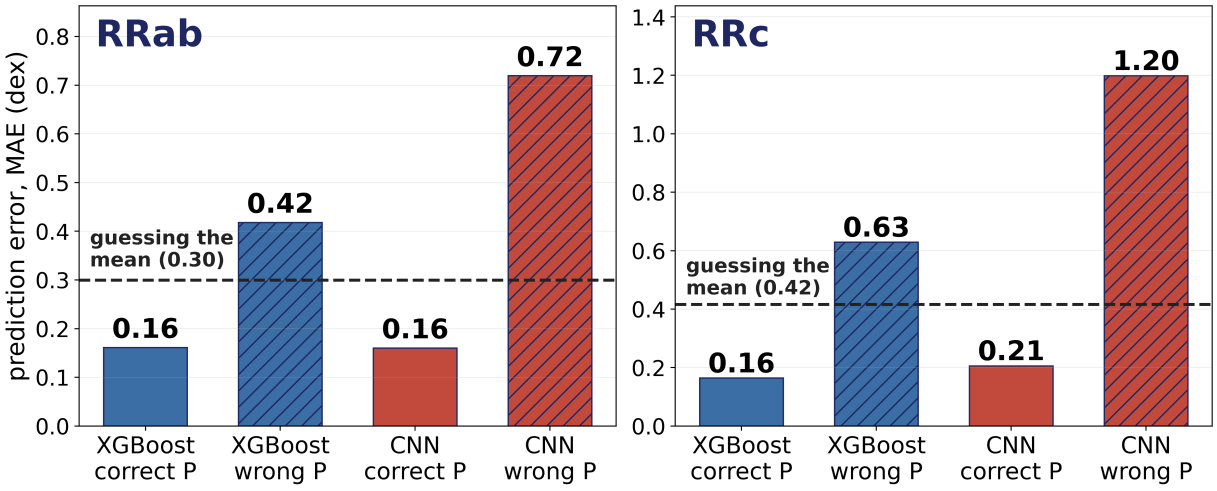


Figure 5 The wrong-period stress test. Solid bars: models evaluated with features at a consistent period; hatched bars: the same models with features computed at a deliberately wrong period. The dashed line is the guess-the-mean baseline for the same stars. A wrong period is worse than no light curve at all, for both model families and both subtypes.

6.2 Recovering Aliased Short-Period RRc

One failure mode proved fully repairable. The original period search used a maximum frequency of 3.5 cycles per day, implicitly flooring the search at periods of 0.286 days; genuine RRc stars with shorter periods were assigned aliased periods above the floor. Re-running the search with a raised frequency ceiling recovered about 105 RRc stars whose folded curves had previously been scrambled (Figure 6); after recovery they occupy the expected short-period region of the period–amplitude (Bailey) diagram and rejoin the usable sample.

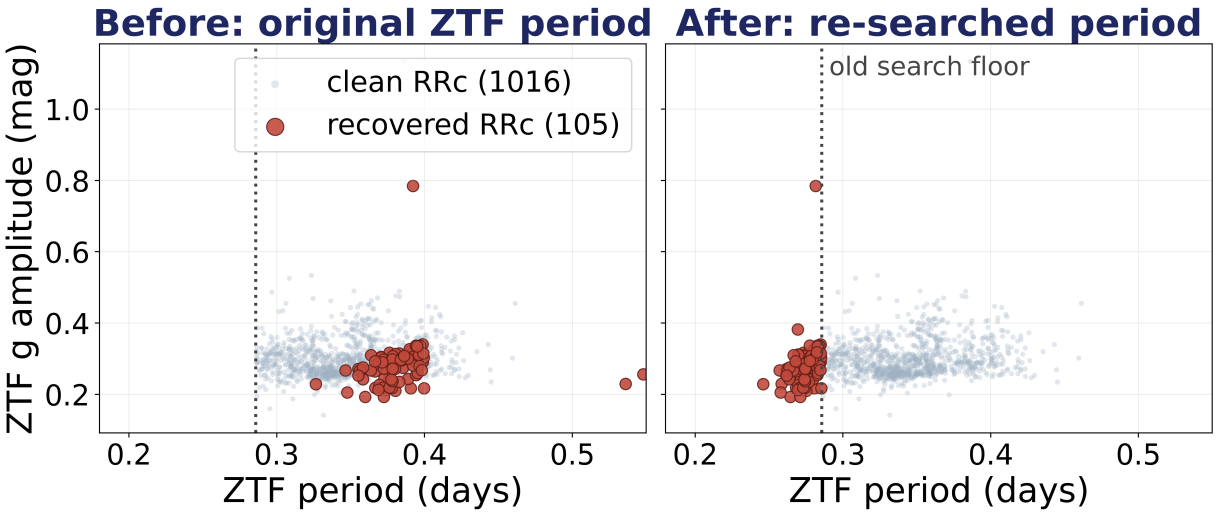


Figure 6 Recovered short-period RRc in the period–amplitude plane. Left: with the original catalog periods the recovered stars (red) sit at aliased periods above the old search floor (dotted line). Right: after the wider period search they move to their true short periods below the floor, consistent with the known RRc population.

6.3 A Period-Reliability Classifier

Not every unreliable period can be repaired, so the remaining cases must at least be flagged. Using only survey-internal quantities (periodogram and light-curve statistics available for every star, with no external cross-match required), we trained a classifier to predict whether a star's period is trustworthy. It separates reliable from unreliable periods with ROC-AUC 0.99 for RRab and 0.92 for RRc, allowing the modeling sample of Section 3 to be restricted to reliable-period stars and providing a reusable quality flag for the production catalog. A recurring lesson of the project, worth stating plainly: reliable preprocessing improved the predictions more than any change of model family.

7 The Catalog: 135,873 Stars, All-Sky

The ZTF-based features cover only part of the sky, so for the full-catalog application we retrain the same recipe on the light-curve parameters that Gaia DR3 publishes for every RR Lyrae [3]: the pulsation period, the G-band peak-to-peak amplitude, and the low-order Fourier shape parameters (φ_{21}, φ_{31}, R_{21}, R_{31}), i.e. the same physics read from a different survey's photometry. Per-subtype models are tuned with cross-validation on the training split only and reach held-out errors of 0.18 dex (RRab, $R^2 = 0.59$) and 0.17 dex (RRc, $R^2 = 0.79$), essentially matching the ZTF-feature models while extending coverage to the whole sky.

Of the 271,779 RR Lyrae in Gaia DR3, 136,049 have the full parameter set published (Gaia omits the Fourier parameters for poorly observed stars); excluding double-mode pulsators, the catalog contains predictions for 135,873 stars (115,410 RRab and 20,463 RRc), roughly 33× the spectroscopic training sample. Each entry carries flags recording whether the star was in the training set and whether it lies inside the training feature domain, along with the per-star uncertainty intervals of Section 8.

Figure 7 shows the catalog on the sky. It reproduces the expected large-scale chemical structure of the Galaxy, with the metal-rich disk concentrated at low galactic latitude and the metal-poor halo dominating away from the plane; known substructures such as the Sagittarius stream stand out against the field. None of this structure was imposed: it emerges from per-star predictions made independently of position.

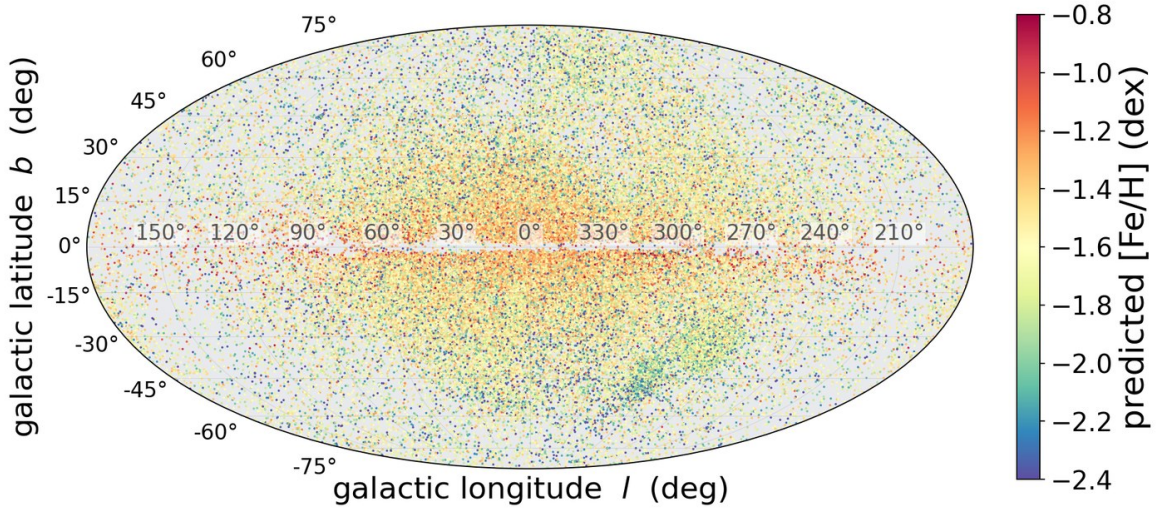


Figure 7 The photometric metallicity catalog: 135,873 RR Lyrae in galactic coordinates (longitude l, latitude b; the Milky Way disk lies at $b \approx 0°$), colored by predicted [Fe/H].

8 External Validation

A held-out test set, however carefully protected, answers a narrower question than users of a catalog will ask, because it is drawn from the same survey, the same label source, and the same distribution as the training data. The catalog was therefore tested where it could genuinely fail: against astronomical systems whose metallicities come from an entirely independent measurement tradition and never entered training at any stage.

8.1 Globular Clusters and Dwarf Galaxies

Stars in the same globular cluster share essentially a single, carefully measured metallicity, so the median prediction over a cluster's RR Lyrae members provides a clean external test. Members are identified with curated membership catalogs built from Gaia astrometry (membership probability > 0.9; cf. [5]), retaining systems with at least three RR Lyrae members; predictions are compared with literature values on four metallicity scales, adopting Harris [4] as the primary one. The same procedure is applied to dwarf spheroidal galaxies using published system metallicities, and Gaia's own published photometric metallicity column is evaluated identically to provide a like-for-like baseline.

Table 2 and Figure 8 summarize the outcome. Across 36 globular clusters and 9 dwarf galaxies, the predicted system medians agree with the literature at approximately the 0.15 dex level (MAE 0.13–0.19 dex depending on the reference scale, with a systematic offset of about −0.11 dex on the primary scale). Evaluated identically, Gaia's published photometric metallicities show 0.21–0.30 dex offsets on the same systems, i.e. roughly twice ours on the primary scale. The dwarf galaxies show a similar systematic offset with larger scatter, dominated by the smallest systems, where few members and genuine internal metallicity spread limit the comparison. We emphasize that no cluster or dwarf value was ever a training label: this is a test outside the training distribution, and the model passes it.

Table 2 Cluster-level validation: MAE and bias (dex) of median member predictions against literature cluster metallicities, per reference scale, compared with Gaia's published photometric metallicity column evaluated identically.

		This work		**Gaia DR3 column**	
Reference scale	N_{clusters}	MAE	bias	MAE	bias
Harris (primary)	36	**0.149**	−0.108	0.298	+0.292
K19	31	0.192	−0.185	0.206	+0.182
B19	24	0.194	−0.155	0.242	+0.228
V20	27	**0.129**	−0.114	0.238	+0.236

8.2 Auditing the Per-Star Uncertainties

Each catalog entry carries an uncertainty interval built from the model's measured mistakes rather than a parametric error model: out-of-fold residuals on the training portion are binned by predicted [Fe/H], and each star inherits the 16th–84th percentile interval of its bin. These intervals were then audited on 449 cluster member stars excluded from training (Figure 9). The raw intervals achieve 64% empirical coverage against a 68% target, and the pull distribution reveals the structure of the miscalibration: the interval *widths* are correct (normalized scatter 0.87, and only 0.61 within a single cluster), while the *centers* are displaced by a global +0.12 dex

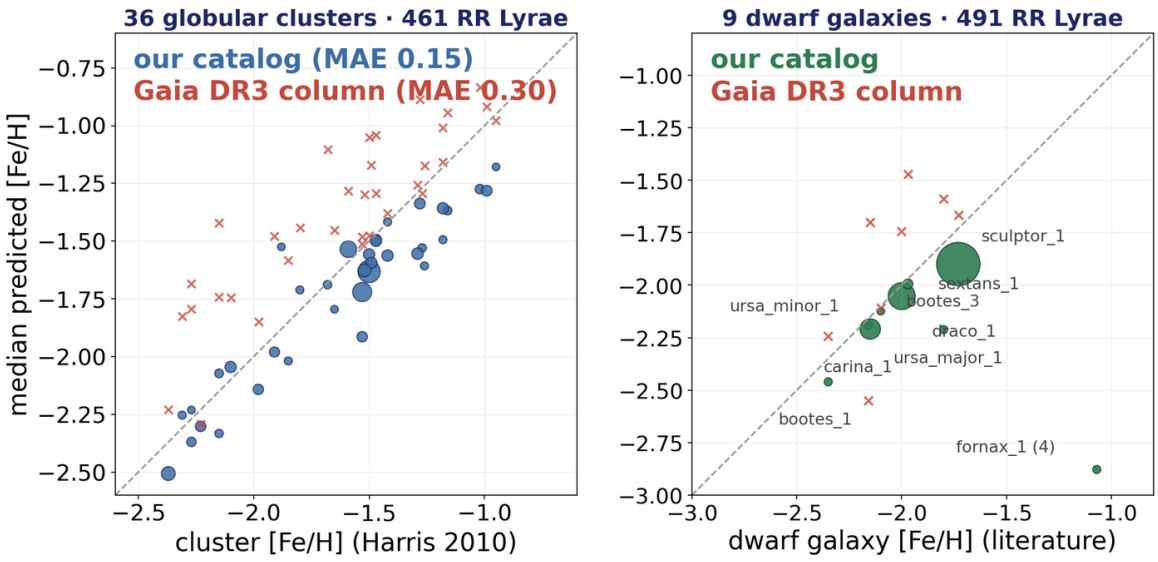


Figure 8 External validation. Each point is one system: literature [Fe/H] (horizontal) against the median prediction for its RR Lyrae members (vertical); the diagonal is perfect agreement. Left: globular clusters. Right: dwarf spheroidal galaxies. Crosses show Gaia's published photometric metallicities for the same systems.

offset between the DESI label scale and the cluster literature scale. Recentering by this single constant restores coverage to 71% overall, while inflating the widths instead makes coverage worse, confirming that the issue is centering rather than size. The one regime where no interval is honest is the metal-rich bin (53% coverage even after recentering), consistent with the information limit of Section 5.3; these stars are flagged in the catalog rather than assigned an error bar we cannot defend. Cluster-to-cluster systematics contribute about 0.14 dex.

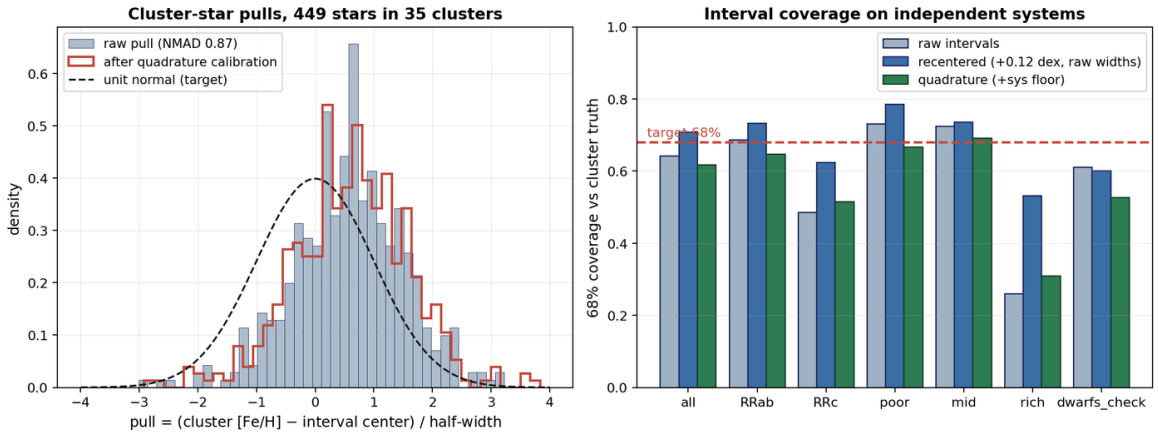


Figure 9 Auditing the per-star intervals on cluster ground truth. Left: the pull distribution (cluster truth minus interval center, in units of the half-width) has approximately the correct width but is shifted, reflecting the DESI-versus-cluster scale offset. Right: empirical 68% coverage by subsample; recentering by +0.12 dex (with unchanged widths) restores near-nominal coverage everywhere except the metal-rich bin.

Key Insight

Trust in a catalog is built outside its training distribution. The cluster audit showed our error bars were the right size but on the wrong ruler, a 0.12 dex scale offset invisible to any internal test; fixing it is a convention choice (which metallicity scale the catalog should be tied to) rather than a modeling task.

9 Self-Supervised Pretraining: A Measured Negative Result

The labeled sample is small, but the unlabeled corpus is not: could representation learning on the full unlabeled population improve on the engineered features? To answer this decisively rather than anecdotally, we assembled the complete Gaia DR3 epoch photometry for the RR Lyrae sample (a bulk download of ~67 GB, yielding usable multi-band light curves for 269,621 stars, roughly 70× the labeled sample) and pretrained recurrent encoders on it with three self-supervised objectives: contrastive learning on augmented views of the same star, masked reconstruction (a quarter of the points hidden and reconstructed), and next-point forecasting. Frozen embeddings were evaluated with regression probes on the DESI-labeled subset against a predict-the-mean null, at three corpus scales (7k, 20k, and the full ~270k), and finally in a hybrid test appending the embeddings to the engineered features over the standard 15 random splits. Pretraining used GPU resources of the Digital Research Alliance of Canada.

Table 3 summarizes the outcome. Scaling the contrastive setup from 7k to the full corpus produced no gain: its probe hovers just below the null at every scale. The pretext objective matters more than the corpus size: masked reconstruction and especially forecasting produce embeddings with substantially more metallicity signal (0.268 versus null 0.302 for RRab; 0.339 versus 0.384 for RRc), consistent with periodicity being the carrier of the chemical information. But the decisive test is the hybrid: appended to the engineered features, the best embeddings leave the 15-split MAE unchanged or slightly worse (0.164 → 0.167 RRab, 0.188 → 0.195 RRc, improving at most 2 of 15 splits). Everything the encoders learn about metallicity, the period and shape features already encode. The representations are genuinely informative in absolute terms, yet redundant at the margin; we report this as a clean negative result, and we consider the measurement itself (three objectives, three scales, and a controlled hybrid test) a useful contribution, since it tells the next project where not to spend its compute at this label count.

Table 3 Self-supervised pretraining on the full Gaia corpus (269,621 stars). Probe values are ridge-regression MAE (dex) on frozen embeddings for the DESI-labeled subset; the null predicts the training mean. The hybrid row appends the forecasting embeddings to the engineered features (15-split mean).

Setting	RRab	RRc
Null (predict the mean)	0.302	0.384
Contrastive embeddings (probe)	0.297	0.361
Masked-reconstruction embeddings (probe)	0.279	0.343
Forecasting embeddings (probe)	**0.268**	**0.339**
Engineered features (reference)	0.164	0.188
Engineered features + embeddings	0.167	0.195

10 Limitations

- **Metallicity scale.** The catalog inherits the DESI label scale, which sits ≈0.12 dex from the cluster literature scale; the per-star interval widths are validated, but the catalog as a whole shifts with the choice of ruler. Adopting a scale is a convention decision we intend to make in consultation with the RR Lyrae community.
- **Metal-rich regime.** Above [Fe/H] ≈ −1.3 the training labels are too few to support either unbiased predictions or honest intervals, and the mitigation study indicates this is not fixable by modeling alone. The affected stars are flagged rather than silently reported.
- **Unresolved period pathologies.** The reliability classifier flags, but cannot repair, the remaining unreliable periods, and some period disagreements may reflect blended neighboring stars rather than aliases; a star-by-star audit against Gaia neighbor counts is in progress with our collaborators.
- **Label noise.** The headline MAE includes the spectroscopic labels' own measurement noise, so the model's intrinsic error is somewhat better than quoted; we have not attempted to deconvolve the two.
- **Representation learning.** Frozen-embedding approaches are now conclusively evaluated, but end-to-end fine-tuning of a pretrained encoder with labels remains untested and could in principle behave differently.

11 Conclusions and Future Work

This internship project has gone from a feasibility question, whether light curves can substitute for spectra, to a delivered and validated data product. Engineered light-curve shape features with gradient-boosted trees, trained on a few thousand DESI labels and protected by a strict evaluation protocol, roughly halve the error of a guess-the-mean baseline; a data-quality investigation identified unreliable periods as the dominant error source and addressed it with a wider period search and a reliability classifier; and the resulting pipeline, applied at Gaia scale, produced photometric metallicities for 135,873 stars. The catalog was then tested where it could genuinely fail, against globular clusters and dwarf galaxies it had never seen, and agrees with the literature at approximately the 0.15 dex level, roughly twice as close as the Gaia mission's own photometric metallicities on the same systems. Each star carries an uncertainty audited against that same ground truth, and the two honest negative results, the metal-rich information limit and the redundancy of self-supervised embeddings, mark out where further modeling effort would and would not pay off.

> **Takeaway**
>
> Domain-informed light-curve features turn a few thousand spectroscopic labels into photometric metallicities for ∼135,000 stars, with 0.16 dex held-out test error and ≈0.15 dex agreement on systems the model never saw. Reliable preprocessing improved the catalog more than any change of model family.

Looking forward, several directions continue beyond the internship:

- completing the journal manuscript consolidating these results, in preparation with my supervisors;
- handing over the catalog, code, and documentation to the research group in a reproducible

form;

- resolving the metallicity-scale convention and incorporating external metal-rich spectroscopy as additional training labels;
- the one untested deep-learning avenue, end-to-end fine-tuning of a pretrained encoder, for which the scaled-up self-supervised runs provide the starting checkpoints; and
- preparing for Rubin/LSST, where the pipeline's sensitivity to period quality makes early survey cadence the critical test case: we plan to quantify period recovery and feature stability on LSST-cadence photometry and to deliver the period-reliability classifier as a reusable tool for the community.

Acknowledgments

This work was carried out during the SUDS (Summer Undergraduate Data Science) program at the Data Sciences Institute, University of Toronto, with the sponsorship of KAUST Academy, King Abdullah University of Science and Technology, to which I am grateful for making my participation possible. I thank my supervisors, Prof. Ting Li and Dr. Gustavo Medina, for their guidance throughout the project, and our collaborators Nina Hernitschek and Sergey Koposov for the ZTF light-curve feature-extraction pipeline, the cluster and dwarf-galaxy membership catalogs, and many helpful discussions. The self-supervised pretraining computations were performed on the Narval cluster of the Digital Research Alliance of Canada.

References

[1] J. Jurcsik and G. Kovács. "Determination of [Fe/H] from the light curves of RR Lyrae stars". In: *Astronomy and Astrophysics* 312 (1996), pp. 111–120.

[2] István Dékány, Eva K. Grebel, and Grzegorz Pojmański. "Metallicity Estimation of RR Lyrae Stars from Their I-band Light Curves". In: *The Astrophysical Journal* 920.1 (2021), p. 33. DOI: 10.3847/1538-4357/ac106f.

[3] G. Clementini, V. Ripepi, A. Garofalo, et al. "Gaia Data Release 3: Specific processing and validation of all-sky RR Lyrae and Cepheid stars – The RR Lyrae sample". In: *Astronomy and Astrophysics* 674 (2023), A18. DOI: 10.1051/0004-6361/202243964.

[4] William E. Harris. "A Catalog of Parameters for Globular Clusters in the Milky Way". In: *The Astronomical Journal* 112 (1996). 2010 edition, arXiv:1012.3224, p. 1487.

[5] Eugene Vasiliev and Holger Baumgardt. "Gaia EDR3 view on galactic globular clusters". In: *Monthly Notices of the Royal Astronomical Society* 505.4 (2021), pp. 5978–6002. DOI: 10.1093/mnras/stab1475.

[6] Eric C. Bellm, Shrinivas R. Kulkarni, Matthew J. Graham, et al. "The Zwicky Transient Facility: System Overview, Performance, and First Results". In: *Publications of the Astronomical Society of the Pacific* 131.995 (2019), p. 018002. DOI: 10.1088/1538-3873/aaecbe.

[7] DESI Collaboration. *Data Release 1 of the Dark Energy Spectroscopic Instrument*. arXiv:2503.14745. https://arxiv.org/abs/2503.14745. 2025.