

Using machine learning to detect outliers of red clump stars

Andrew Meng¹, Emil Khan², Arun Arunachalam², Shyamal Mitra³

¹ Clements High School, Sugar Land, Texas

² Westlake High School, Austin, Texas

³ The University of Texas at Austin, Austin, Texas

SUMMARY

Red clump stars (RCs) are a subset of red giants with relatively constant luminosities, a unique property that allows astronomers to use them as distance indicators in astrophysical research. Accurate distance measurements help astronomers reduce errors in mapping the Galactic Bar, minimizing parallax errors and depths of the Small Magellanic Galaxy. While literature regarding RCs suggests they exhibit constant luminosity, recent observations suggest luminosity variability that is remarkably correlated with age and metallicity of RCs. This uncertainty makes it crucial to identify the outlier group of RCs and determine which stellar properties are most associated with this change in luminosity. We hypothesized that age would be more strongly associated with luminosity variability than metallicity. We used two machine learning models, Extreme Gradient Boosting (XGB) and Isolation Forest (IF), to evaluate feature importance and identify outliers, respectively. The XGB model showed that age had the highest feature importance (72.6%) in predicting luminosity variability, whereas metallicity had a substantially lower feature importance (9.24%). The IF model successfully identified RCs with unusual luminosity or age values as outliers and detected 74.89% of artificially added outliers. Our findings demonstrate that machine learning successfully detected variability in luminosity, which aligned with our hypothesis. Our results will help astronomers understand the influence of stellar parameters on luminosity variability while accurately classifying variability in RCs luminosity to reduce computational inaccuracies in distance estimates.

INTRODUCTION

We live in a three-dimensional universe, yet the objects on the celestial sphere can only be perceived in two dimensions (1). As a result, the universe seems to have no depth, causing the distance to become an unknown variable for each night sky object relative to each other (2). Without accurate distance measurements, astronomers cannot reliably determine important stellar properties such as a star's true luminosity or its physical location within a galaxy (3). Distance errors also distort measurements of the spatial distribution of stars, making it more difficult to reconstruct the three-dimensional structure of the Milky Way and understand its formation and evolution (4). One method astronomers use to obtain reliable stellar distance measurements is through standard candles, stellar objects with known luminosities that allow distances to

be estimated from their observed brightness (5). Red clump stars (RCs) are commonly used as standard candles because of their relatively consistent luminosities (6).

To understand why RCs are useful as distance indicators, it is important to consider the physical properties of these stars. RCs are a subset of red giants located on the cooler horizontal branch of the Hertzsprung-Russell (HR) diagram (7). The HR diagram shows the relationship between a star's luminosity and surface temperature (8). Astronomers have traditionally considered RCs luminosity to be relatively insensitive to differences in stellar age and metallicity, meaning that variation in these properties were expected to produce only small changes in luminosity (9). This unique characteristic has made them standard candles in astrophysical research, allowing astronomers to accurately gauge distances to other stellar objects like star clusters in which they are embedded, as well as derive other physical parameters relative to them (10–12). Because RCs serve as standard candles, they have played an important role in determining the age and size of the Milky Way galaxy (13). RCs have been used to detect the Galactic Bar, minimize parallax errors, and reduce errors in measurements of the depth of the Small Magellanic Galaxy (10, 14). However, recent observations suggest that RCs luminosity varies with stellar age and metallicity. The relative influence of these properties remains unclear, raising concerns about the reliability of RCs as distance indicators (14,15). Detecting and classifying this variability by identifying outliers is therefore important for reducing errors in distance calculations and evaluating the reliability of RCs as standard candles (15).

Large astronomical datasets often contain measurements for tens of thousands of stars, making individual examination impractical (16). Machine learning (ML) algorithms provide an efficient way to detect patterns and sift through large amounts of data quickly (17). Isolation Forest (IF) and Extreme Gradient Boosting (XGB) are two machine learning approaches suited to anomaly detection and regression, respectively. IF is an unsupervised ML model for anomaly detection that repeatedly divides data into smaller groups based on the values of individual input features. Each division creates two possible groups, forming a binary tree structure (18). IF identifies outliers by measuring how quickly each data point becomes separated from the rest of the dataset, with points isolated after fewer divisions considered more unusual (18). The model is useful for large datasets because it identifies anomalies efficiently while requiring relatively little memory (19). IF can be calibrated by changing its hyperparameters, such as the number of estimators and the contamination

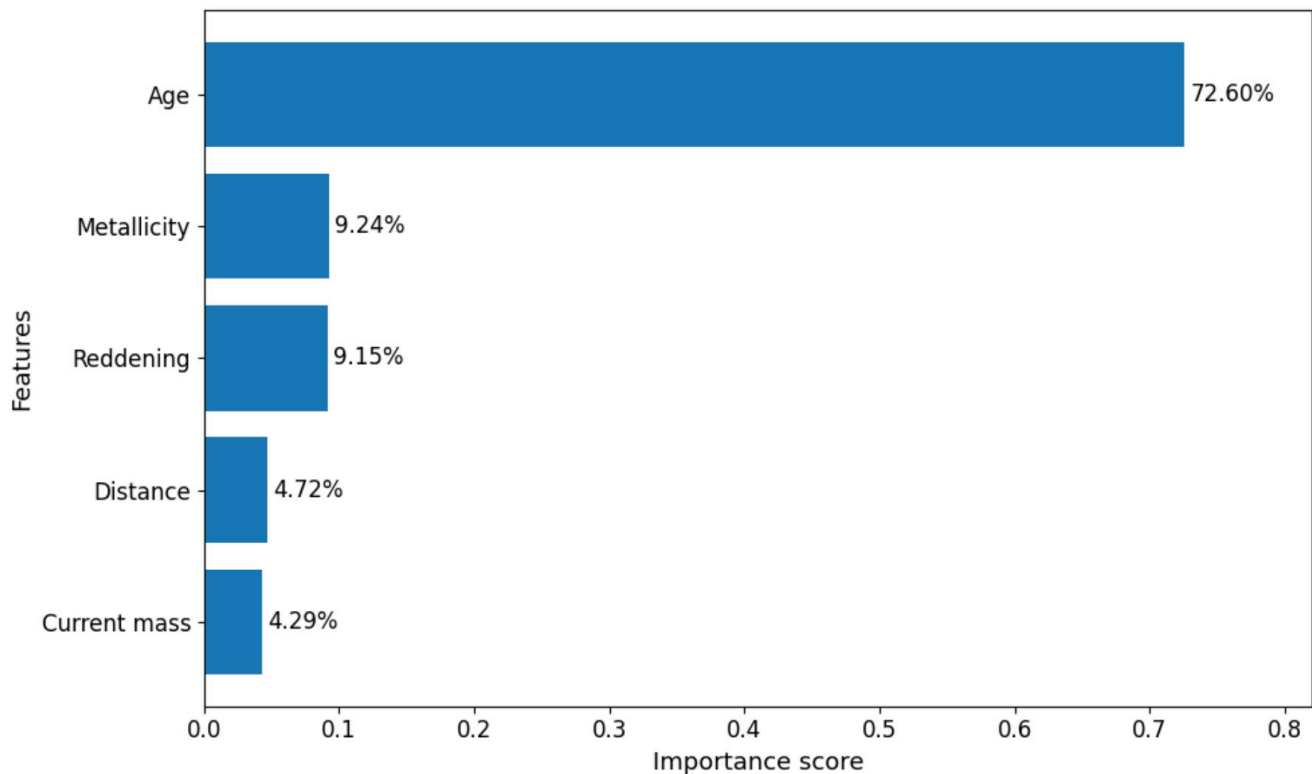


Figure 1. Importance score for each parameter using the Extreme Gradient Boost (XGB) model. Bars show the percentage feature importance scores for age in gigayears, metallicity ([Fe/H]), reddening ($E(B-V)$), distance in kiloparsecs, and current mass in solar masses. The model used a maximum tree depth of 3, a learning rate of 0.1, and 100 decision trees.

value (20). The number of estimators determines the number of isolation trees in the model, while the contamination value specifies the expected proportion of outliers in the dataset (20). XGB, or Extreme Gradient Boosting, is a supervised ML model used for regression and classification that applies ensemble learning and gradient boosting (21). Gradient boosting combines many weak models to form a more robust model that minimizes prediction error. XGB is useful for predicting values and evaluating the relative importance of input features. XGB can be calibrated by changing its learning rate, which determines the pace at which the algorithm learns, and the max depth, which limits the number of divisions each decision tree can make (21).

Applying IF and XGB to RCs data provides a way to identify stars with unusual luminosity and determine which stellar properties contribute most strongly to luminosity variation. More accurately identifying these outliers and establishing the key properties associated with luminosity variation provides a route toward improving the reliability of RCs-based distance measurements (10,14). We hypothesized that RCs luminosity varies by an extensive range of stellar properties, and that stellar age would be more strongly associated with this variability than metallicity. We analyzed 3,978 RCs from the GALAH survey using IF and XGB (16). IF was used to identify RCs with unusual luminosity and stellar age values, while XGB was used to predict luminosity and evaluate feature importance. Age showed a substantially higher feature importance than metallicity, and IF identified a group of outlier RCs. Our results supported our hypothesis and showed that ML methods were useful for examining

stellar properties associated with RCs luminosity variability. Improving the reliability of RCs as distance indicators will support more accurate studies of the structure and evolution of galaxies.

RESULTS

This study examined how strongly several stellar properties were associated with RCs luminosity variability, with particular emphasis on stellar age and metallicity. We also investigated whether RCs with unusual age or luminosity values formed a distinct outlier group. To address these questions, we used XGB to predict luminosity and evaluate feature importance, IF to identify outliers, and statistical analyses to compare the relationships of age and metallicity with luminosity.

XGB Prediction and Feature Analysis

To evaluate the importance of key stellar properties on luminosity, we trained the XGB model on the GALAH RCs dataset using five stellar properties – age, metallicity, reddening, distance, and mass – as input features. The feature importance analysis showed that age had the highest importance on luminosity at 72.6%, followed by metallicity at 9.24%, reddening at 9.15%, distance at 4.72%, and mass at 4.29% (**Figure 1**). The model produced an R^2 value of 0.907 on the testing dataset, indicating a strong fit between the predicted and observed log luminosity values (**Figure 2**). To examine how prediction error changed across log luminosity values, we calculated the residuals between the predicted and observed log luminosities and grouped them into intervals of 0.5 (**Figure 3**). For log luminosity values between

2 and 2.5, the mean residual was 0.000328, whereas it was -0.0292 for values between 2.5 and 3. The magnitude of the mean residual was therefore approximately 89 times greater in the higher interval. This suggests a large rise in error as log luminosity increases past 2.5, but no statistical test was conducted to confirm significance (**Figure 3**). A separate linear regression between metallicity and luminosity produced an R^2 score of 0.286. This indicated that metallicity explained only 28.6% of the variation in luminosity, while most of the variation remained unexplained by the model (**Figure 4**).

IF outlier detection

We implemented an IF model to identify outliers based on age and log luminosity (**Figure 5**). We augmented the dataset with 442 artificial outlier points designated as unlikely to belong to the RCs population. The IF model classified the data as either inliers or outliers and correctly identified 331 of the artificial outliers, corresponding to a detection rate of

74.89%. Among the 3,978 original RCs, the model classified 111 stars as outliers, representing 2.79% of the RCs sample.

Analysis revealed that RCs with log luminosity outside the range 1.6–1.9 or with ages less than 1 gigayear (Gyr) or greater than 8 Gyr were generally flagged as outliers (**Figure 5**). Converting the log luminosity values to actual luminosities in Watts (W), the actual luminosity of RCs is between 1.5×10^{28} - 3.05×10^{28} W or 39 – 80 times brighter than the sun.

Statistical Analysis of Age vs. Metallicity Impact on Stellar Luminosity

For our partial correlation analysis, we controlled the mass, distance, and extinction, while only investigating age and metallicity. Our results show that age exhibits a statistically significant negative correlation with luminosity ($r = -0.563$, $p = 0.001$), accounting for approximately 31.6% of the variance ($r^2 = 0.316$) (**Figure 6A**). Metallicity shows no significant relationship ($r = -0.06$, $p = 0.703$), explaining less

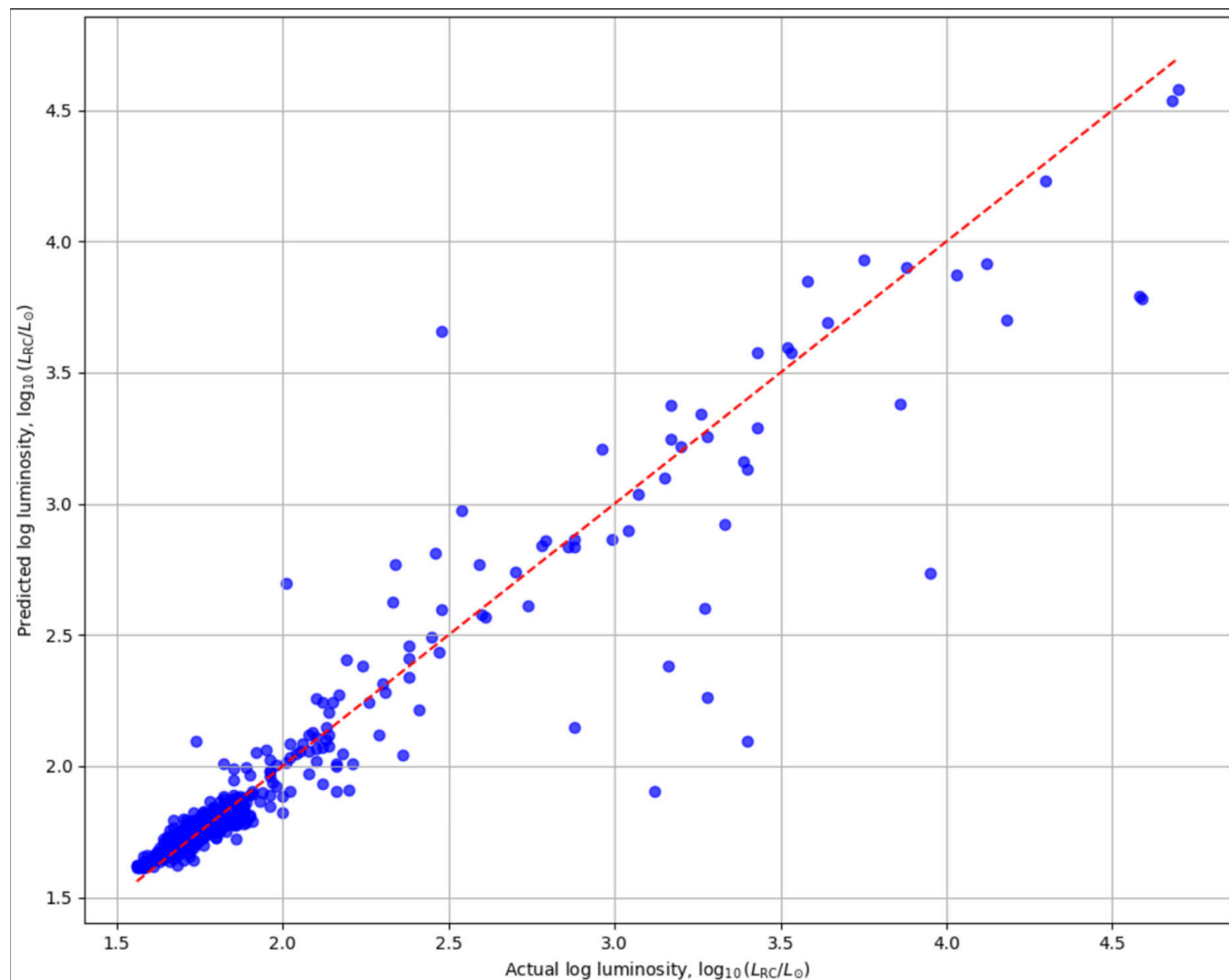


Figure 2. Actual vs. predicted log luminosity for the Extreme Gradient Boost (XGB) model testing set. XGB used 5 stellar input features: age, metallicity, reddening, distance, and current mass. The model was trained on 80% of the dataset and evaluated on the remaining 20% testing set. The dashed red line represents perfect agreement between actual and predicted log luminosity values. The testing set R^2 was 0.907. LRC represents red clump stars luminosity, and L_{RC} represents solar luminosity.

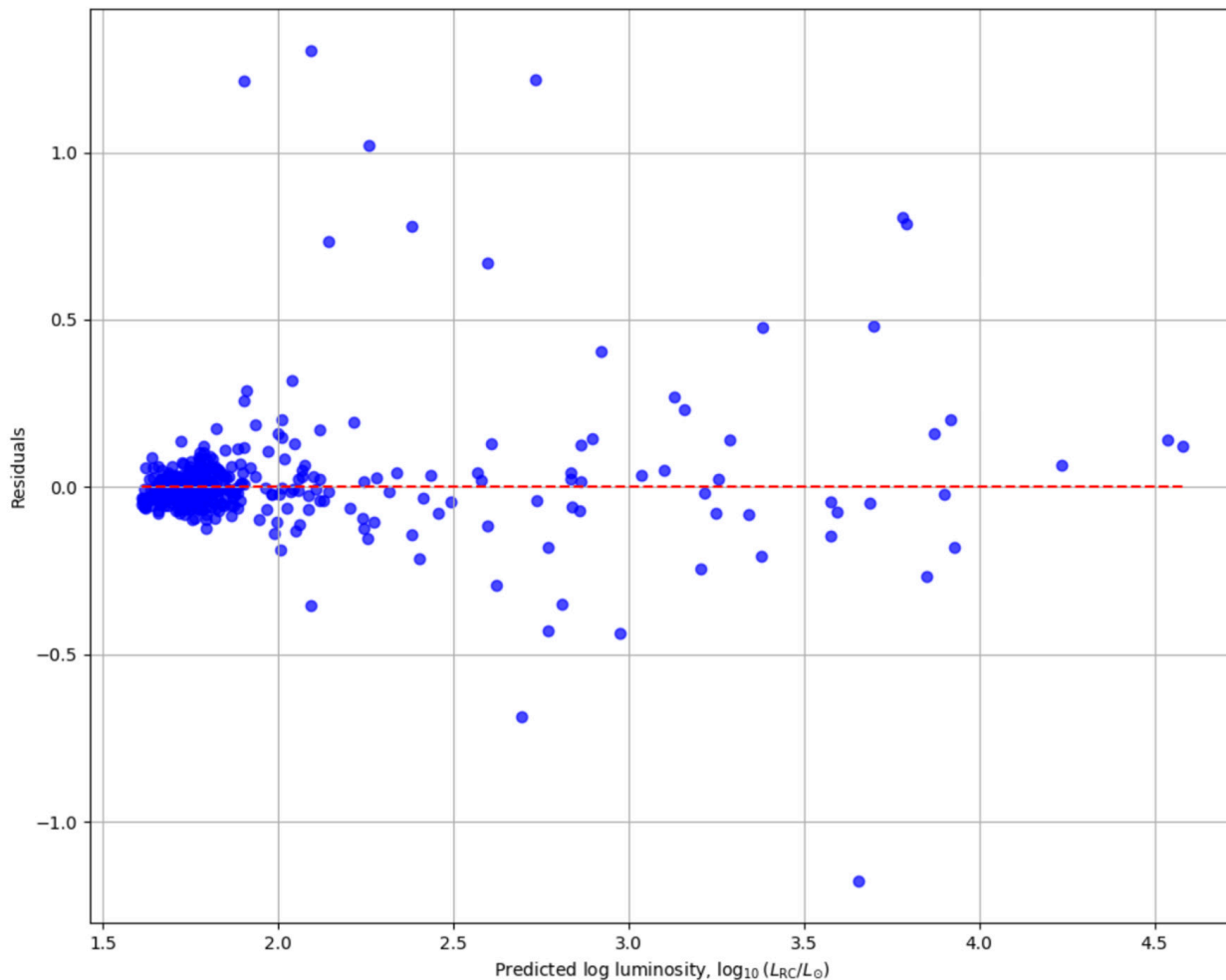


Figure 3. Residual plot using the Extreme Gradient Boost (XGB) model. A graph comparing the predicted log luminosity with the residual from the actual luminosity. The dashed red line marks a residual of zero and serves as a reference for the magnitude of prediction errors. L_{RC} represents red clump stars luminosity, and $L_{\odot}$ represents solar luminosity.

than 0.4% of luminosity variation (**Figure 6B**). We conducted independent Welch's t-tests where we used median splits within age and metallicity to compare luminosity distributions. We created 2 groups for both age and metallicity, where the median value for age was 0.42 Gyr, and metallicity was -1.66 [Fe/H]. Our results indicate that age has a higher impact on luminosity while metallicity has a significantly lower impact. RCs younger than the median age (4.2 Gyr) are more luminous than older stars ($p < 0.0001$, Cohen's $d = 2.48$) (**Figure 7A**). Meanwhile, RCs with high or low metallicity groups show no significant luminosity difference ($p = 0.162$, Cohen's $d = 0.57$) (**Figure 7B**). The Cohen's d of age is 4.35x larger than that of metallicity, demonstrating that age is significantly more determinant of stellar luminosity in our sample.

DISCUSSION

RCs are commonly used as standard candles for astronomical distance measurements yet uncertainty about how stellar age and metallicity are associated with luminosity variability may reduce the accuracy of these estimates. This study evaluated whether stellar age or metallicity was more strongly associated with RCs luminosity variability

and whether ML models could identify unusual RCs in the GALAH dataset. The XGB ML model showed that age had the highest feature importance on RCs luminosity, while metallicity had substantially lower importance. Notably, XGB demonstrated strong predictive performance for RCs luminosity, achieving an R^2 of 0.907 on the unseen testing dataset. Partial correlation analysis supported this difference, as age showed a significant negative relationship with luminosity after controlling for mass, distance, and reddening, whereas metallicity showed no significant relationship. A second ML model, IF, also detected 74.89% of the artificial outliers and identified a small subset of observed RCs with unusual age or luminosity values. These ML models analyzed multiple stellar properties across thousands of RCs while learning predictive relationships and identifying unusual observations. Their strong performance across these tasks demonstrates the effectiveness of ML for studying complex patterns in RCs luminosity variability. Together, these results support our hypothesis that age was a stronger predictor of RCs luminosity variability than metallicity and show that ML could identify a subset of outlier RCs.

Using the XGB model, our work concluded that metallicity

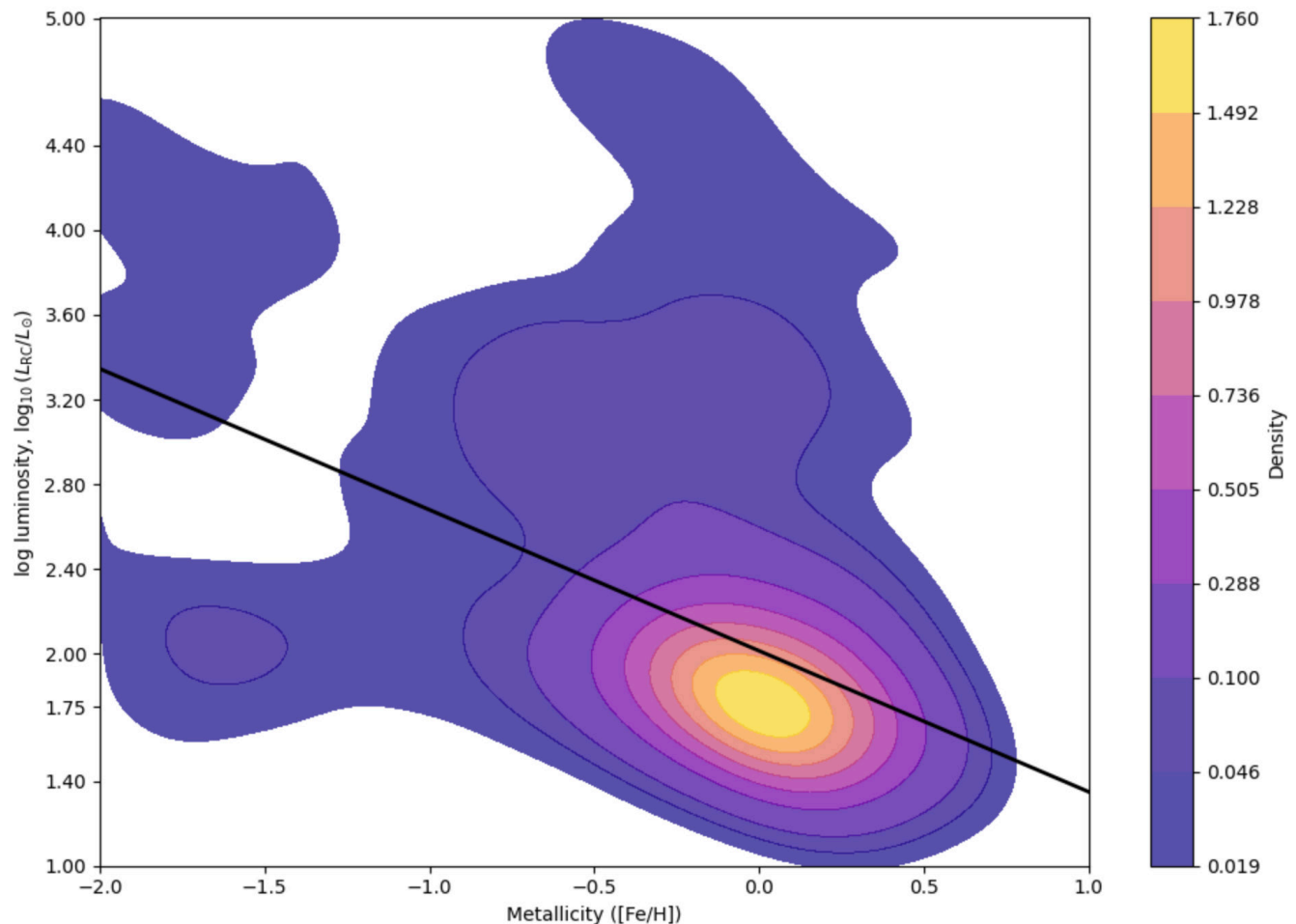


Figure 4. Heatmap of metallicity vs. log luminosity. The color scale represents the estimated local density of RCs, with purple indicating lower density and yellow indicating higher density. Contour lines connect regions of equal density. The black line shows the linear regression fit, with $R^2 = 0.286$. L_{RC} represents red clump stars luminosity, and $L_{\odot}$ represents solar luminosity.

only weakly determines luminosity variability, as shown by a low parameter importance score of 9.24%. Previous work have shown similar results without the use of ML by using photometric observations of Galactic star clusters to compare RCs luminosity across different ages and metallicities, finding a relatively weak dependence on metallicity (14). Another study observed 284 RCs manually and similarly found only a weak relationship between metallicity and luminosity (22). Although theoretical models may show a strong relationship between metallicity and luminosity, they do not account for all sources of luminosity variation (23). One theoretical work considered additional properties, including helium abundance and core mass, and found that these factors also influence RCs luminosity (24). This suggests that some luminosity variation attributed to metallicity may instead reflect the combined effect of other stellar properties. This finding is consistent with the difference between our linear regression and partial correlation results for metallicity and luminosity. The decrease from $R^2 = 0.286$ in the linear regression to $r^2 = 0.004$ in the partial correlation suggests that much of the apparent metallicity-luminosity relationship was associated with mass, distance, and reddening. Thus, metallicity alone has a limited influence on RCs luminosity variability. The

relative stability of RCs luminosity across a range of metallicity values supports the use of RCs as standard candles.

In contrast, age had the highest XGB feature importance at 72.6% and remained moderately associated with luminosity in the partial correlation analysis, with $r^2 = 0.316$ and $p = 0.001$. The greater association between age and luminosity may be related to stellar evolutionary stage. In a shell hydrogen-burning red giant, luminosity increases as energy production shifts to the shell while the core becomes degenerate (25). Once the star's core reaches a certain temperature, helium fusion begins rapidly in an event known as the helium flash. Because the degenerate core cannot initially expand and cool in response to the rising temperature, helium fusion briefly accelerates. The energy released removes the core's degeneracy, causing the core to expand and cool, while the star's outer layers contract (25). As the star becomes less luminous and hotter, it moves from the red giant branch to the horizontal branch on the HR diagram. This transition, known as the helium flash, marks the beginning of stable helium fusion in the star's core (25). This age-dependent transition might help explain the nearly vertical concentration of points at young ages in the IF plot of log luminosity as a function of age. Most RCs younger than 1 Gyr within this concentration

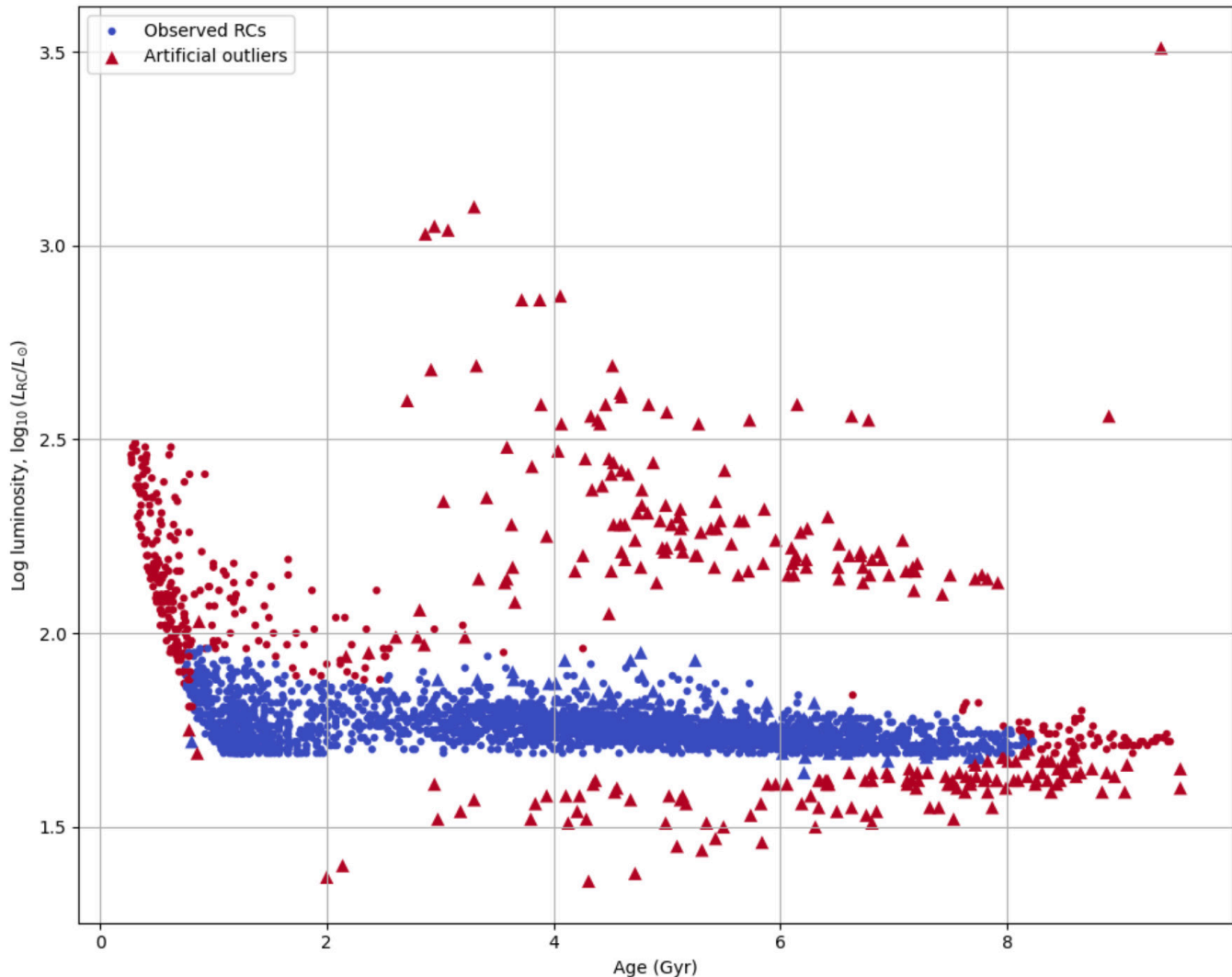


Figure 5. Detecting outliers using the Isolation Forest model. The plot shows log luminosity versus age in gigayears for 3,978 red clump stars (circle markers) and 442 artificially added outliers (triangle markers), comprising 10% of the total sample. The model classifies the data into two classes: blue points represent inliers, while red points represent outliers. L_{RC} represents red clump stars luminosity, and $L_{\odot}$ represents solar luminosity.

were classified as outliers, which may reflect younger stars still transitioning into the horizontal branch. Similarly, most RCs older than 8 Gyr were classified as outliers, possibly reflecting older stars transitioning out of the horizontal branch. These evolutionary differences may help explain the luminosity variability observed among the RCs identified as outliers.

In this study, several factors and limitations may have influenced our results. Although the GALAH dataset contained 3,978 RCs, increasing the data size or finding more data from other data sources would provide more confidence and consistency in our results. Both models' hyperparameters could be further optimized to avoid overfitting the data. In addition, the parameter representing the probability that an observation is a RCs, is a model-derived estimate that may be subject to uncertainty. It is vital to question whether the outliers could still be RCs and whether future studies are needed to better understand the range of constant luminosity for RCs. Future research may aim to implement other

machine learning and deep learning models in more diverse and extensive datasets to validate our findings further.

While our analysis demonstrates a statistically robust relationship between age and luminosity ($p < 0.001$, $d = 2.48$), we acknowledge two important considerations. Age may correlate with unmeasured evolutionary parameters or other confounding factors, such as stellar rotation evolution (26). Furthermore, stellar age estimates carry greater uncertainty than more directly observed properties like mass (27). The metallicity analysis also has a potential limitation because $[\text{Fe}/\text{H}]$ measures iron abundance rather than a star's complete chemical composition (28). Other chemical properties, such as helium or alpha-element abundance, may contribute to luminosity variation that $[\text{Fe}/\text{H}]$ alone does not capture (29). Therefore, our results identify associations between these stellar properties and luminosity but do not establish direct causal relationships. Future work incorporating more measured evolutionary parameters and chemical properties could help address these limitations.

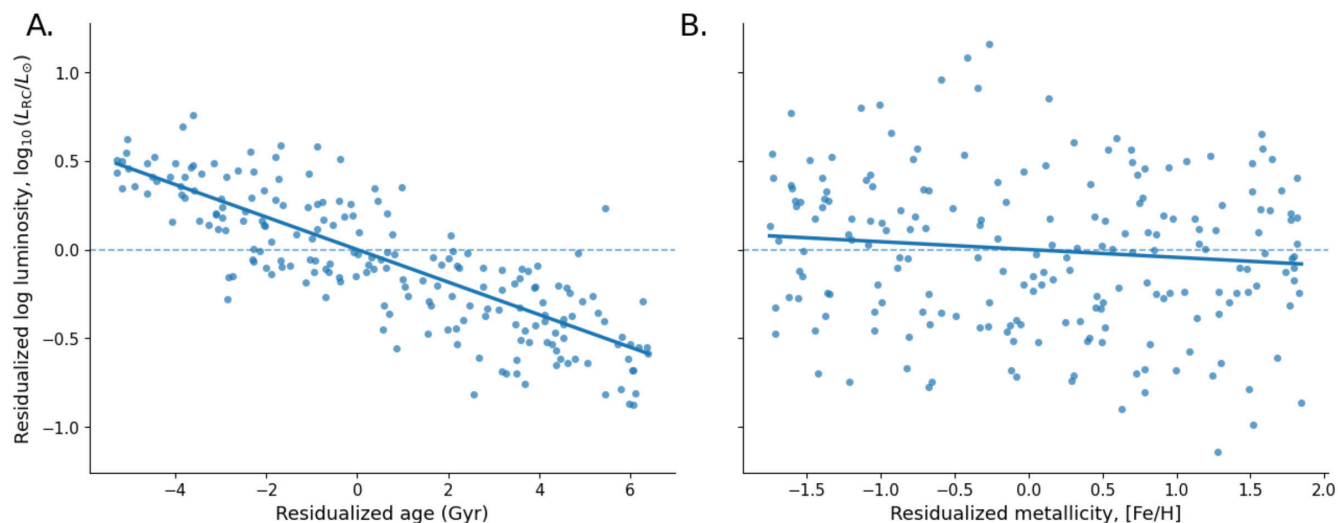


Figure 6. Partial correlation of red clump star (RCs) log luminosity with age and metallicity. Residualized values represent the variation remaining after removing the linear effects of current mass, distance, and reddening. (A) Residualized log luminosity plotted against residualized age, measured in gigayears. Age showed a significant negative partial correlation with log luminosity, with $r = -0.563$ and $p = 0.001$. (B) Residualized log luminosity plotted against residualized metallicity ($[\text{Fe}/\text{H}]$). Metallicity showed no significant partial correlation with log luminosity, with $r = -0.060$ and $p = 0.703$. Points represent individual RCs, solid lines represent linear regression fits, and dashed horizontal lines mark a luminosity residual of zero. $[\text{Fe}/\text{H}]$ represents the logarithmic iron abundance relative to hydrogen compared with the Sun. L_{RC} represents red clump stars luminosity, and $L_{\odot}$ represents solar luminosity.

Our study found that age had a high impact and metallicity had a relatively low impact on luminosity in our dataset, while also demonstrating the effectiveness of ML models in detecting RCs luminosity variability. The high predictive performance of the XGB model and the effective outlier detection by the IF model underscores the ever-increasing potential of ML in astrophysics. Future studies using larger datasets, additional stellar properties, and further model optimization could validate these findings and improve outlier detection. These approaches may help astronomers assess the reliability of RCs as standard candles, identify potential

outliers, and reduce errors in distance calculations crucial to understanding the universe's structure.

MATERIALS AND METHODS

Software Libraries and Environment

In our study, we utilized Jupyter Notebook to write and execute all the codes and models created using Python (30). The versatile software libraries used in implementing the models were Numpy, Pandas, and Scikit_Learn (31). Matplotlib and Seaborn were used to create visual graphs (32). The two machine learning models used were XGBoost,

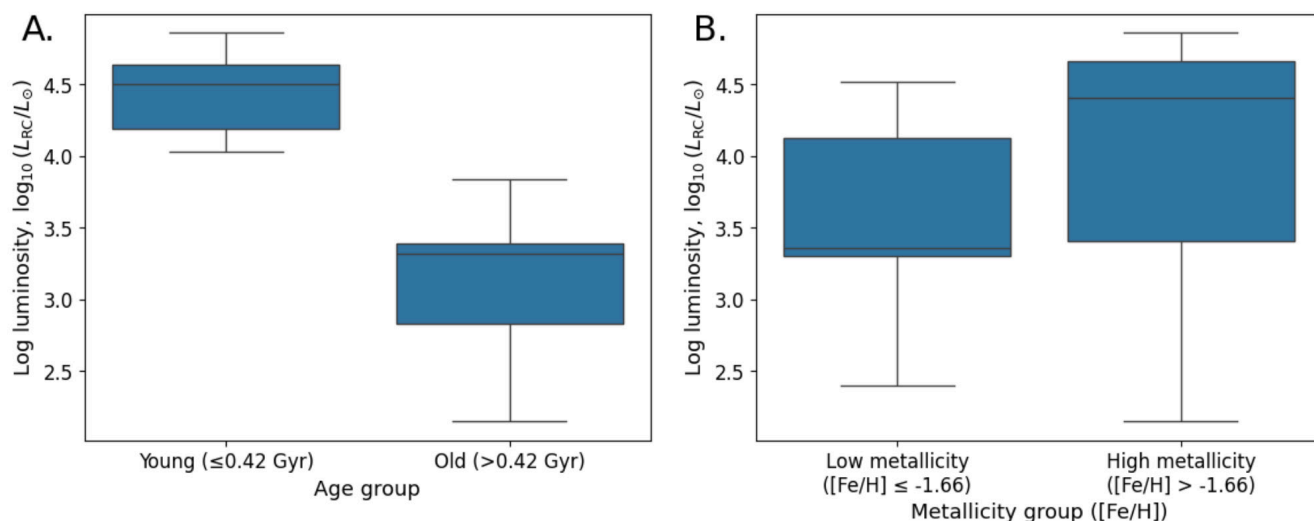


Figure 7. Distribution of red clump star log luminosity by age and metallicity. (A) Boxplot comparing younger stars, with ages below 4.2 gigayears (Gyr), and older stars, with ages above 4.2 Gyr. The groups differed significantly (Welch's t-test: $p = 5.1 \times 10^{-10}$, $d = 2.48$). (B) Boxplot comparing stars with lower metallicity, $[\text{Fe}/\text{H}] < -1.66$, and higher metallicity, $[\text{Fe}/\text{H}] > -1.66$. The groups did not differ significantly ($p = 0.162$, $d = 0.57$). $[\text{Fe}/\text{H}]$ represents the logarithmic iron abundance relative to hydrogen compared with the Sun. L_{RC} represents red clump stars luminosity, and $L_{\odot}$ represents solar luminosity.

which we imported from the XGBoost library, and the IF Model, which we imported from the Scikit_Learn ensemble library (21,24).

Data Acquisition and Preprocessing

The dataset for this study was collected from the GALAH Survey, which contained 52 parameters of around 203,851 stars (60% red giants) (17). It included essential parameters such as age, log luminosity, metallicity, reddening, distance, mass, and the percentage chance the star is a red clump (*is_redclump*) (18). Log luminosity (*l*) is defined as the base-10 logarithm of a star's luminosity (L_{RC}) relative to the sun (L_{sun}),

$$l = \log_{10} \left(\frac{L_{RC}}{L_{sun}} \right) \quad (\text{Eq. 1})$$

which is used because stellar luminosities span several orders of magnitude, making comparisons more manageable. To recover the actual luminosity of RCs in watts, we use:

$$L_{RC} = L_{sun} \cdot 10^l \quad (\text{Eq. 2})$$

where $L_{sun} = 3.828 \times 10^{26}$ W is the solar luminosity (33).

Metallicity ($[Fe/H]$) represents the logarithmic ratio of a star's iron (Fe) abundance relative to hydrogen (H), compared to that of the Sun:

$$\left[\frac{Fe}{H} \right] = \log_{10} \left(\frac{N_{Fe}}{N_H} \right)_{star} - \log_{10} \left(\frac{N_{Fe}}{N_H} \right)_{sun} \quad (\text{Eq. 3})$$

where N_{Fe} and N_H are the number densities of Fe and H, respectively. This value is important because it serves as a proxy for the star's overall chemical composition, which can influence its evolution and luminosity (34).

Reddening ($E(B-V)$) is defined as the difference between the observed and intrinsic color indices of a star:

$$E(B-V) = (B-V)_{observed} - (B-V)_{intrinsic} \quad (\text{Eq. 4})$$

where $(B-V)_{observed}$ is the measured color index of the star, which includes the effects of interstellar dust, and $(B-V)_{intrinsic}$ is the star's true color index, defined as the difference between its intrinsic blue (*B*) and visual (*V*) magnitudes in the absence of interstellar dust. Reddening causes distant stars to appear redder than closer stars due to the scattering of blue light from interstellar dust. Knowing the $E(B-V)$ value allows prediction of the spectral class, which can be used to estimate the star's temperature and distance (35).

Our data preprocessing, included filtering out stars that did not have an *is_redclump* of 100%, removing irrelevant parameters for determining log luminosity, and deleting any rows containing null values for any of the critical parameters. After preprocessing, only 3,978 RCs remained, and all features were normalized to ensure consistent scaling. Our newly processed data set was split into 80% training and 20% testing sets for future supervised learning using the XGBoost model.

XGBoost (XGB) Machine Learning Model

The XGB model was utilized to predict log luminosity after training on the input parameters identified during preprocessing. We gave it five functional parameters in

determining luminosity: age, metallicity, distance, mass, and reddening. The hyperparameters we chose were *max_depth*, *learning_rate*, and *n_estimators*. We set *max_depth* to its default value of 3, *n_estimators* to the default of 100, and *learning_rate* to 0.1 (24). Although the default *learning_rate* is 0.3, we considered it relatively high and potentially prone to overfitting, so we opted for a lower value. The model was trained using the processed dataset, and its effectiveness was evaluated through the R^2 score, which measures the proportion of variance explained by the model (36) (Figure 2).

Isolation Forest (IF) Machine Learning Model

The IF model was utilized to detect outliers from the features. The number of estimators determines the number of isolation trees in the forest, while the contamination value is the threshold for the point to become an anomaly. This model was crucial for our study as we wanted to detect and group the outliers based on their parameters. We used the default hyperparameters provided by the library, with number of estimators set to 100 and contamination value set to 0.1 (10% of the data points are classified as outliers) (19, 20). These default settings are generally robust and provide reliable outlier detection without requiring formal tuning. This model was trained on the relevant features to classify RCs and identify potential outliers that could skew distance calculations. The model's effectiveness was evaluated through multiple passes and supplemented with the theory that RCs have consistent luminosity. This helped verify that outliers were present and that the model could accurately detect these to help with distance calculation and the integrity of the red clump's standard candle property (10, 13, 16).

Statistical Methods

Partial correlation was chosen to evaluate the relationship between each predictor (age or metallicity) and stellar luminosity while controlling for potential confounding variables such as mass, distance, and reddening. This approach isolates the direct effect of each variable, allowing us to determine which factors independently contribute to luminosity variation (37).

Welch's t-test was applied to determine whether stars in different groups of the same variable (age or metallicity) show significant differences in luminosity. By splitting the sample into two groups at the median of each variable, we could directly compare whether "younger vs. older" stars or "low-metallicity vs. high-metallicity" stars have systematically different luminosities (Figure 6). This test is ideal because it does not assume equal variances between groups, which is often the case in astrophysical data. Using Welch's t-test allowed us to quantify and test the significance of these group-level differences in luminosity (38). Results with $p < 0.05$ were considered statistically significant for both analyses. Together, these methods allow us to both isolate individual effects and test for meaningful differences between groups, giving a robust statistical understanding of the factors driving stellar luminosity.

ACKNOWLEDGMENTS

We thank our peer mentors, Juhi Malwade, Hanyu Wei, and Anthony Yang, for their feedback and support. We also

thank the High School Research Academy program for its pivotal role in enabling us to conduct this research. Lastly, we thank Dr. Gebhardt for his insight and valuable suggestions for improving our project.

REFERENCES

1. "How Do I See Depth?" *SCECinfo*. <https://scecinfo.usc.edu/geowall/stereohow.html>. Accessed 7 July 2024.
2. "Depth of Space." Griffith Observatory. 29 Dec. 2021. <https://griffithobservatory.org/exhibits/gunther-depths-of-space/depth-of-space/>. Accessed 7 July 2024.
3. González-Santamaría, I., et al. "Properties of Central Stars of Planetary Nebulae with Distances in Gaia DR2." *Astronomy & Astrophysics*, vol. 630, Oct. 2019, A150. <https://doi.org/10.1051/0004-6361/201936162>.
4. "Gaia Sees Strange Stars in Most Detailed Milky Way Survey to Date." *European Space Agency*. 13 June 2022. https://www.esa.int/Newsroom/Press_Releases/Gaia_sees_strange_stars_in_most_detailed_Milky_Way_survey_to_date. Accessed 9 Aug. 2026.
5. "Caldwell 51." *NASA Science*. 23 Aug. 2024. <https://science.nasa.gov/mission/hubble/science/explore-the-night-sky/hubble-caldwell-catalog/caldwell-51/>. Accessed 9 Aug. 2026.
6. Khanna, Shourya, et al. "The GALAH Survey: Velocity Fluctuations in the Milky Way Using Red Clump Giants." *Monthly Notices of the Royal Astronomical Society*, vol. 482, no. 3, Jan. 2019, pp. 4215–4232. <https://doi.org/10.1093/mnras/sty2924>.
7. Girardi, Léo. "Red Clump Stars." *Annual Review of Astronomy and Astrophysics*, vol. 54, Sept. 2016, pp. 95–133. <https://doi.org/10.1146/annurev-astro-081915-023354>.
8. Fraknoi, Andrew, et al. "The H–R Diagram." *Astronomy 2e*, OpenStax, 2022. <https://openstax.org/books/astronomy-2e/pages/18-4-the-h-r-diagram>. Accessed 25 July 2026.
9. Plewa, Philipp. "Gaia and the Red Clump." *Astrobites*. 30 Nov. 2017. <https://astrobites.org/2017/11/30/gaia-and-the-red-clump/>. Accessed 7 July 2024.
10. Paczyński, Bohdan and Krzysztof Z. Stanek. "Galactocentric Distance with the OGLE and Hipparcos Red Clump Stars." *The Astrophysical Journal*, vol. 494, no. 2, Feb. 1998, pp. L219–L222. <https://doi.org/10.1086/311181>.
11. "Cosmology Standard Candle Not So Standard After All." *NASA Jet Propulsion Laboratory*. <https://www.jpl.nasa.gov/news/cosmology-standard-candle-not-so-standard-after-all/>. Accessed 9 July 2024.
12. He, Xu-Jiang, et al. "Identification, Mass, and Age of Primary Red Clump Stars from Spectral Features Derived with the LAMOST DR7." *Monthly Notices of the Royal Astronomical Society*, vol. 512, no. 2, 21 Mar. 2022, pp. 1710–1721. <https://doi.org/10.1093/mnras/stac484>.
13. "Measuring the Distance to the Far Side of the Galaxy." *Center for Astrophysics | Harvard & Smithsonian*. <https://pweb.cfa.harvard.edu/news/measuring-distance-far-side-galaxy>. Accessed 10 July 2024.
14. Sarajedini, Ata. "WIYN Open Cluster Study. III. The Observed Variation of the Red Clump Luminosity and Color with Metallicity and Age." *The Astronomical Journal*, vol. 118, no. 5, Nov. 1999, pp. 2321–2326. <https://doi.org/10.1086/301112>.
15. Seidel, Edward, et al. "The Evolution of Red Clump Stars: Theoretical Sequences." *The Astrophysical Journal Supplement Series*, vol. 63, 1 Apr. 1987, pp. 917–945. <https://doi.org/10.1086/191186>.
16. "GALAH DR3." *GALAH Survey*. <https://www.galah-survey.org/dr3/overview/>. Accessed 16 July 2024.
17. "Machine Learning." *Center for Astrophysics | Harvard & Smithsonian*. 2024. <https://pweb.cfa.harvard.edu/research/topic/machine-learning>. Accessed 16 July 2024.
18. "IsolationForest." *scikit-learn*. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html>. Accessed 19 July 2024.
19. Cortes, David. "An Introduction to Isolation Forests." *CRAN*. https://cran.r-project.org/web/packages/isotree/vignettes/An_Introduction_to_Isolation_Forests.html. Accessed 20 July 2024.
20. Dhiraj K. "Anomaly Detection Using Isolation Forest in Python." *Paperspace by DigitalOcean Blog*. 20 Aug. 2024. <https://blog.paperspace.com/anomaly-detection-isolation-forest/>. Accessed 9 Aug. 2026.
21. Adrian. "Tuning Parameters for Gradient Boosting/XGBoost." *Data Science Stack Exchange*. 19 May 2020. <https://datascience.stackexchange.com/questions/74488/tuning-parameters-for-gradient-boosting-xgboost>. Accessed 15 Sept. 2025.
22. Udalski, Andrzej. "The Optical Gravitational Lensing Experiment: Red Clump Stars as a Distance Indicator." *The Astrophysical Journal*, vol. 531, no. 1, Mar. 2000, pp. L25–L28. <https://doi.org/10.1086/312513>.
23. Girardi, Léo and Maurizio Salaris. "Population Effects on the Red Giant Clump Absolute Magnitude, and Distance Determinations to Nearby Galaxies." *Monthly Notices of the Royal Astronomical Society*, vol. 323, no. 1, May 2001, pp. 109–129. <https://doi.org/10.1046/j.1365-8711.2001.04084.x>.
24. Cole, Andrew A. "Age, Metallicity, and the Distance to the Magellanic Clouds from Red Clump Stars." *The Astrophysical Journal*, vol. 500, no. 2, 20 June 1998, pp. L137–L140. <https://doi.org/10.1086/311422>.
25. "Post-Main-Sequence Sun." *University of Nevada, Las Vegas*. https://www.physics.unlv.edu/~jeffery/astro/star/post_main_sequence/post_main_sequence_sun.html. Accessed 12 Apr. 2025.
26. Gruner, David, et al. "New Insights into the Rotational Evolution of Near-Solar-Age Stars from the Open Cluster M 67." *Astronomy & Astrophysics*, vol. 672, Apr. 2023, article A159. <https://doi.org/10.1051/0004-6361/202345942>.
27. Barber, Madyson G. and Andrew W. Mann. "Using the Gaia Excess Uncertainty as a Proxy for Stellar Variability and Age." *The Astrophysical Journal*, vol. 953, no. 2, Aug. 2023, article 127. <https://doi.org/10.3847/1538-4357/ace044>.
28. Mowlavi, Nami, et al. "Stellar Mass and Age Determinations. I. Grids of Stellar Models from $Z = 0.006$ to 0.04 and $M = 0.5$ to $3.5 M_{\odot}$." *Astronomy & Astrophysics*, vol. 541, May 2012, A41. <https://doi.org/10.1051/0004-6361/201117749>.
29. Joo, Seok-Joo, et al. "New Insight on the Origin of the Double Red Clump in the Milky Way Bulge." *The Astrophysical Journal*, vol. 840, no. 2, May 2017, p. 98. <https://doi.org/10.3847/1538-4357/aa6d4f>.
30. Kluyver, Thomas, et al. "Jupyter Notebooks—A Publish-

- ing Format for Reproducible Computational Workflows.” Positioning and Power in Academic Publishing: Players, Agents and Agendas, edited by Fernando Loizides and Birgit Schmidt, IOS Press, 2016, pp. 87–90. <https://doi.org/10.3233/978-1-61499-649-1-87>.
31. Pedregosa, Fabian, *et al.* “Scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825–2830. <https://www.jmlr.org/papers/v12/pedregosa11a.html>. Accessed 28 July 2026.
 32. Hunter, John D. “Matplotlib: A 2D Graphics Environment.” *Computing in Science & Engineering*, vol. 9, no. 3, May–June 2007, pp. 90–95. <https://doi.org/10.1109/MCSE.2007.55>.
 33. Prša, Andrej, *et al.* “Nominal Values for Selected Solar and Planetary Quantities: IAU 2015 Resolution B3.” *The Astronomical Journal*, vol. 152, no. 2, Aug. 2016, article 41. <https://doi.org/10.3847/0004-6256/152/2/41>.
 34. Fraknoi, Andrew, *et al.* “Using Spectra to Measure Stellar Radius, Composition, and Motion.” *Astronomy 2e*, OpenStax, 2022. <https://openstax.org/books/astronomy-2e/pages/17-4-using-spectra-to-measure-stellar-radius-composition-and-motion>. Accessed 28 July 2026.
 35. Fraknoi, Andrew, *et al.* “Colors of Stars.” *Astronomy 2e*, OpenStax, 2022. <https://openstax.org/books/astronomy-2e/pages/17-2-colors-of-stars>. Accessed 28 July 2026.
 36. Fernando, Jason. “R-Squared: Definition, Formula, Uses, and Limitations.” *Investopedia*. <https://www.investopedia.com/terms/r/r-squared.asp>. Accessed 8 Aug. 2024.
 37. “Partial Correlation Using SPSS Statistics.” *Laerd Statistics*. <https://statistics.laerd.com/spss-tutorials/partial-correlation-using-spss-statistics.php>. Accessed 7 Jan. 2026.
 38. Albright, Elizabeth A. “Two Independent Samples Unequal Variance (Welch’s Test).” *ENV710 Statistics Review Website*, Nicholas School of the Environment, Duke University. <https://sites.nicholas.duke.edu/statsreview/means/welch/>. Accessed 7 Jan. 2026.

Received: November 16, 2024

Accepted: February 02, 2026

Published: October 10, 2026

Copyright: © 2026 Meng, Khan, Arunachalam and Mitra. All JEI articles are distributed under the Creative Commons Attribution Noncommercial No Derivatives 4.0 International License. This means that you are free to share, copy, redistribute, remix, transform, or build upon the material for any purpose, provided that you credit the original author and source, include a link to the license, indicate any changes that were made, and make no representation that JEI or the original author(s) endorse you or your use of the work. The full details of the license are available at <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>.