SCHOLARLY DEBUT
by Scholarly Review

Spring 2026

A Machine Learning Method for Classifying Variable Stars

By Shreenidhi Anand and Sahasra Lagudu

AUTHOR BIO

Shreenidhi Anand is a high school student in Texas. She plans to pursue a career in electrical engineering.

Sahasra Lagudu is an incoming freshman at UT Austin. She attended a high school in central Texas and is planning to major in engineering.

ABSTRACT

Classifying variable stars like Cepheids and RR Lyrae variables is a vital part of exploring the universe. Understanding these stars can provide deep insights into measuring cosmic distances through their period-luminosity relationship. Using machine learning, we aim to classify Cepheid Types I, Cepheid Type II, and RR Lyrae Variables based on redshift, color indices, galactic coordinates, and metallicity. Our initial step was to gather a dataset of approximately 6000 stars for RR Lyrae and 2300 for Cepheids and split them into 2 classes: Cepheid and RR Lyrae from the SIMBAD database, then randomize our sample to avoid bias and graph the data to identify useful relationships. Our graphs showed differences in metallicity, galactic coordinates, and redshift, which led us to include only these factors in our classification model. Our accuracy on the Random Forest model was 92%, and on the Gradient Boosting model, it was 89%. These results suggest strong links between these parameters and the star types they indicate. This study aims to enhance existing classification models by adding parameters that could improve their accuracy and provide more statistical insights into the information each parameter provides about star types. However, we recognize the limitations of our method, including the lack of hyperparameter tuning and confidence intervals.

Keywords*: Machine learning, Gradient Boosting, Random Forest, Variable stars, Cepheid, RR Lyrae.*

SCHOLARLY DEBUT
by Scholarly Review

Spring 2026

INTRODUCTION

Many studies have used machine learning for stellar classification to develop new methods or improve existing ones. For example, the study by T. Kuntzer, M. Tewes, and F. Courbin explores a machine-learning approach for classifying stars into spectral types using broadband images and diffraction patterns (Kuntzer et al., 2016). Similar to this research, we aim to create a machine-learning model to classify variable stars based on their physical properties rather than light curves. Variable stars and their classification are essential in determining cosmological distances. In space, where depth perception is lacking, one standard method for measuring distance is by using variable stars. Astronomers employ the period-luminosity relationship to compare actual brightness with observed brightness and then apply the inverse square law to estimate distance.

Therefore, a machine learning model was developed to classify variable stars based on properties other than their light curves. Since many studies rely on light curves for classification, other properties were used in our research. Light curves are more typically used as the data is easier to gather, and there is a clear relationship found. Consequently, our study focused on the question: "Is there a correlation between properties such as galactic coordinates, metallicity, redshift, and color indices and specific types of variable stars?" We aim to develop an accurate classification model by identifying relationships between variable stars and various properties.

Although we achieved this goal, our study only included types of variable stars such as Type I Cepheids, Type II Cepheids, and RR Lyrae, as well as stars within our galaxy. As a result, our model is currently only applicable for intergalactic classification, and future research should incorporate extragalactic stars to broaden its practical application. For each feature, we generated plots with the collected data to understand the variance among classification groups. These results helped us identify the most relevant properties to include in the model, and we refined the model to improve accuracy.

METHODOLOGY

This study was primarily quantitative, as we relied on numerical data from the SIMBAD database and conducted numerical analysis to identify patterns among the classification groups. Previous studies on the classification of variable stars (Kim & Coryn A. L. Bailer-Jones, 2015) were referenced to understand the parameters used in their models. However, no direct reference was used to create our model.

We performed queries specific to the SIMBAD database to retrieve data for the necessary parameters for RR Lyrae, Cepheid Type I (Classical Cepheids), and Cepheid Type II stars. Data was collected for the following parameters: Galactic Coordinates, Redshift, Metallicity (Fe / H), and Color Indices (*SIMBAD Astronomical Database - CDS (Strasbourg)*, n.d.). SIMBAD astronomical database. https://simbad.cds.unistra.fr/simbad/) et al., 1990.

Parameter	**Explanation**
Galactic Coordinates	A celestial object's coordinates (galactic longitude, galactic latitude)

	on the celestial sphere relative to the Milky Way galaxy. Galactic Longitude is obtained by measuring the angle between a star and the galactic plane. Galactic Latitude is obtained by measuring the degrees north or south of the celestial equator, a great circle that aligns with the galaxy's plane (*Sky Maps with Pierre Auger Data*, 2025).
Metallicity (Fe / H coefficient)	The amount of heavy metals (heavier than hydrogen and helium) in stars relative to the amount in the sun. Obtained through calibrated instruments and spectrophotometry to determine the abundance of elements in the star's atmosphere (*Metallicity of Stars*, n.d.).
Redshift	Displacement of an astronomical object away from the observer due to a change in wavelength, as observed by the Doppler Effect. Obtained by comparing the expected spectrum of a star to one that was created in a laboratory (Las Cumbres Observatory, 2023).
Color Indices (B-V and V-R)	Determines the color of the star by subtracting two filters. The lower the color index, the bluer the object appears. The magnitude of the star is observed through a color filter, and that magnitude is put on a scale (*Color Indices and Surface Temperature*, n.d.).

Table 1. Definitions of Parameters Used in the Classifier (All already available on the SIMBAD database)

Jupyter Notebook and the Python programming language, along with the libraries Pandas, NumPy, Seaborn, and Matplotlib, were used to perform data analysis and create visualizations to evaluate the suitability of each factor for inclusion in the model. After acquiring and organizing the data, all stars with null distance values were cleaned, and all distances were converted to parsecs. Next, histograms were plotted to identify which stars are within or outside the Milky Way galaxy. The following plots were generated:

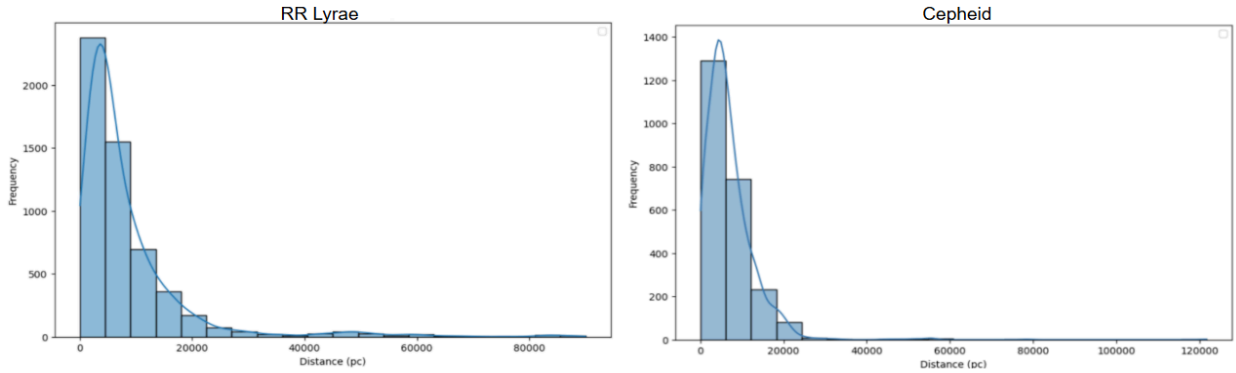

Fig. 3: Histogram of the Distance Distribution of RR Lyrae Variable Stars and Cepheid Variables. This graph helps identify at what distance the band of our galaxy ends. This distance is found to be 30000 parsecs, so stars that have a greater distance were excluded from our study.

A trough occurs at approximately 30,000 parsecs, which helps us differentiate between stars within our galaxy and those outside. Stars within this boundary are considered for our classification model, while those outside are filtered out to ensure the accuracy and relevance of our results.

Because only a few representative stars were found outside the Milky Way, these stars were removed from the dataset. As a result, 5460 RR Lyrae Variables were identified within the galaxy, and 124 were outside it. Additionally, 2355 Cepheids (both Type I and Type II) were found inside the galaxy, while nine were outside. We acknowledge that excluding stars outside the Milky Way could limit the generalizability of our model, as these stars might significantly influence the classification, and we encourage future research to explore this. To analyze this data, we used an 80-20 train-test split ratio. We recognize that there is a bias due to the much higher number of RR Lyrae stars, and this imbalance could have been addressed with resampling methods such as SMOTE (Synthetic Minority Oversampling Technique). Future investigations in this field are likely to benefit from this improvement. Next, the distribution of each parameter within each classification group was examined using various plots.

Metallicity

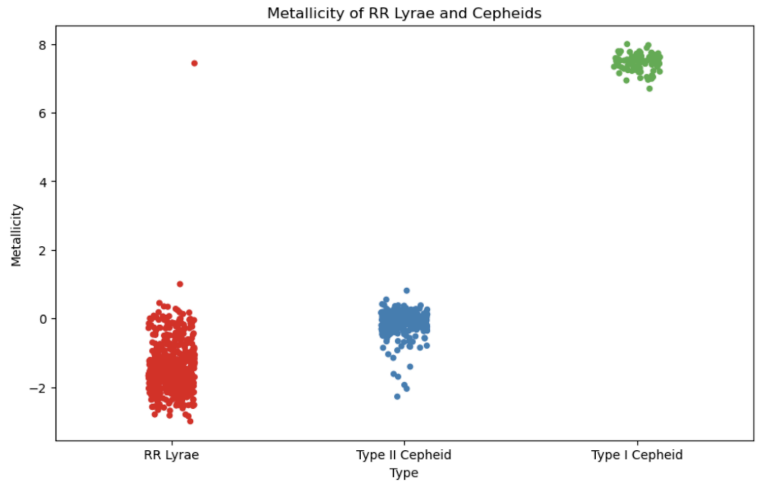

Fig 4: Strip Plot of the Metallicity Distributions for Each Type of Variable Star. Since there is a visible difference, which is that RR Lyrae stars have the lowest metallicity, followed by Cepheid type 2, and Cepheid type 1 stars have the highest metallicity, metallicity was used as a classification parameter.

As hypothesized, RR Lyrae stars tend to be less metal-rich because they are older. There is an outlier with a very high metallicity, which could be a misclassification or an actual outlier. Type II Cepheids also tend to have lower metallicities, although they have a smaller range, and Type I Cepheids are clearly the most metal-rich stars. The distributions are shown by the boxplot below:

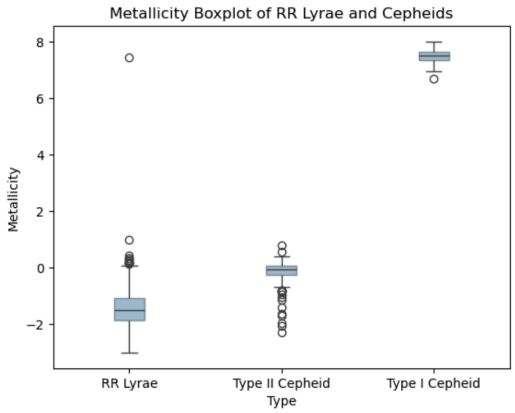

Fig 5: Boxplot showing the Metallicity Distribution for Each Type of Variable Star. This graph highlights the stark difference in metallicity range between RR Lyrae and Cepheid stars, especially between RR Lyrae and Cepheid Type 1 stars, which led us to select it as a classification parameter.

Because the medians of each of the three parameters clearly distinguished each other, and the spreads of each type of star had minimal overlaps, we concluded that this parameter would be useful for classifying these stars.

Color Indices

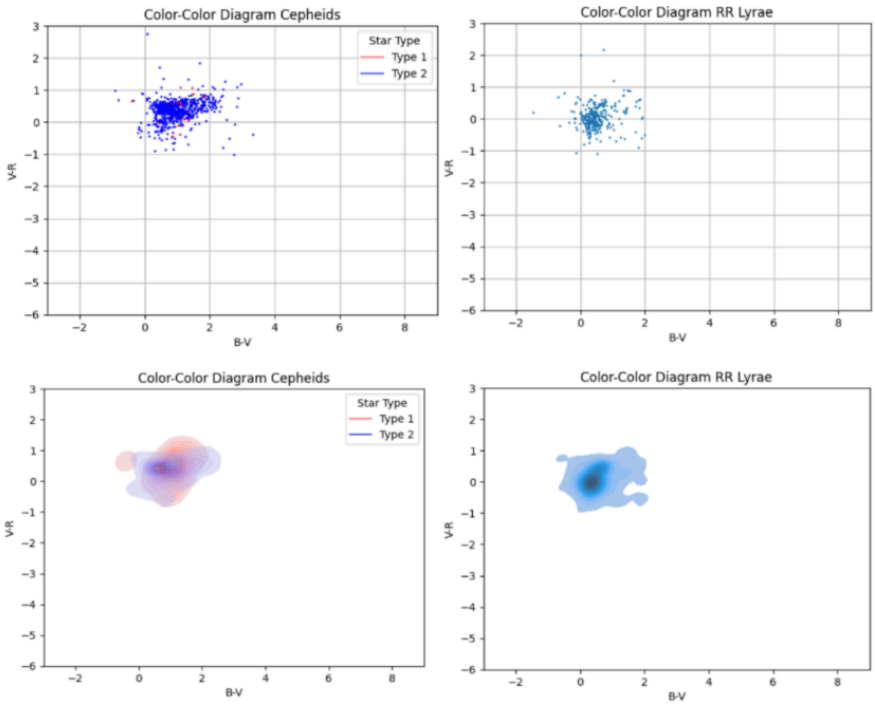

Fig 6: Color-Color Diagrams for Each Type of Variable Star (top), density plots for color-color diagrams (bottom). The graphs illustrate that there isn't enough of a difference between the color-color clusters of the types of stars classified, so this parameter was excluded from the model.

The diagrams above display the color-color plots, which show the relationship between the B-V and V-R indices. These indices were preferred over other filters because they are more commonly available in the SIMBAD database, whereas other color filters had too many missing values to include in the study. As shown in the density plot (Fig. 6), the clustering differences within each classification group are not statistically significant, so the feature was omitted from the model.

Galactic Latitude

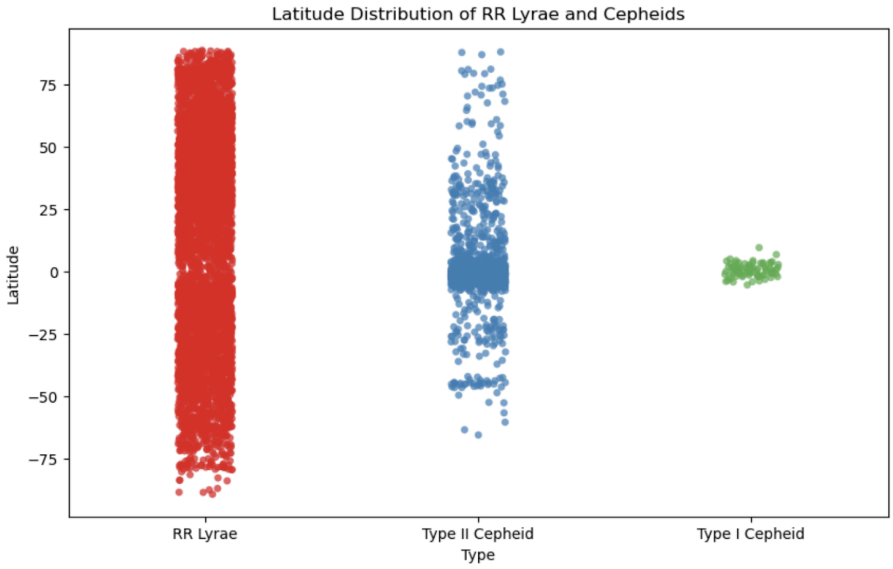

Fig 8: Strip Plot Showing the Galactic Latitude Distributions for Each Type of Variable Star. The difference between the latitude distributions of the three types of variable stars results in this parameter being utilized in our machine learning model.

As shown, RR Lyraes are generally found at all galactic latitudes, while Type I Cepheids tend to be concentrated near the galactic plane. Conversely, Type II Cepheids are strongly clustered near the plane, though several outliers are also present outside this range. The following box plot effectively illustrates this:

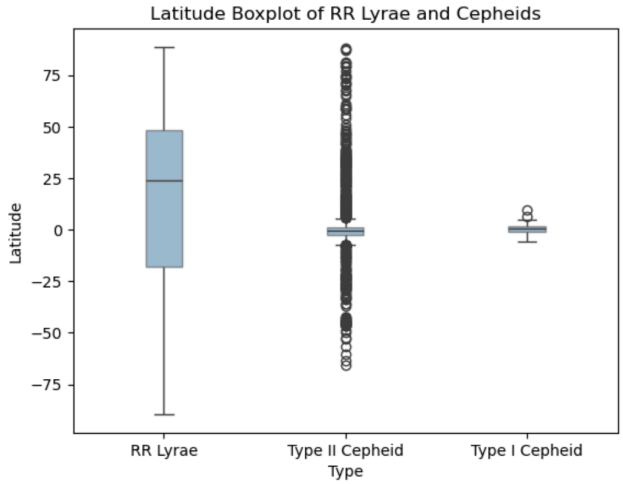

Fig 9: Box Plot Showing the Galactic Latitude Distributions for Each Type of Variable Star. This graph emphasizes the fact that, though there are outliers, the difference in range of the galactic latitude makes it a viable parameter to use.

This feature was added to the classification model because of the general differences in the median across the three stars and the galactic latitude ranges for each classification group.

Redshift

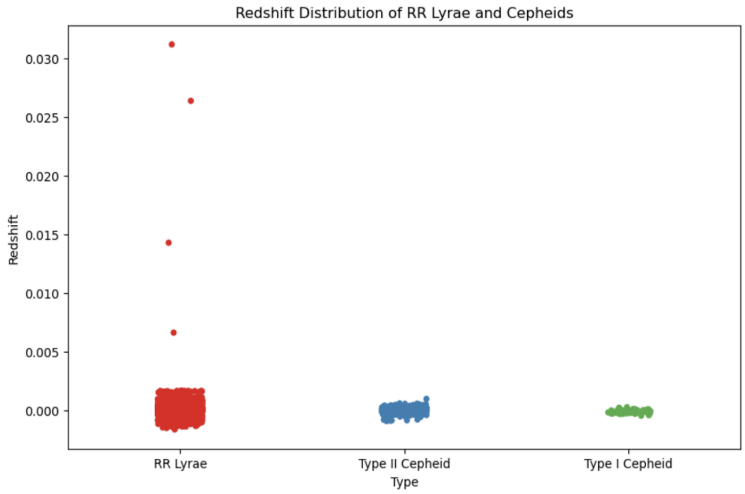

Fig 10: Strip Plot Showing the Redshift Distributions for Each Type of Variable Star. The varying range in the redshift distributions for each type of star emphasizes the value of this parameter in our overall classification methods.

The chart above shows the relationships between the redshifts of Type I (green) and Type II (blue) Cepheids and RR Lyrae (red). Four of the RR Lyrae stars have unusually high redshift values. This may result from measurement errors caused by the instruments. It could also be due to gravitational anomalies or interactions that may have increased their radial velocity, leading to a higher redshift. A noticeable difference is also apparent in the data point ranges between the Cepheid and RR Lyrae variables.

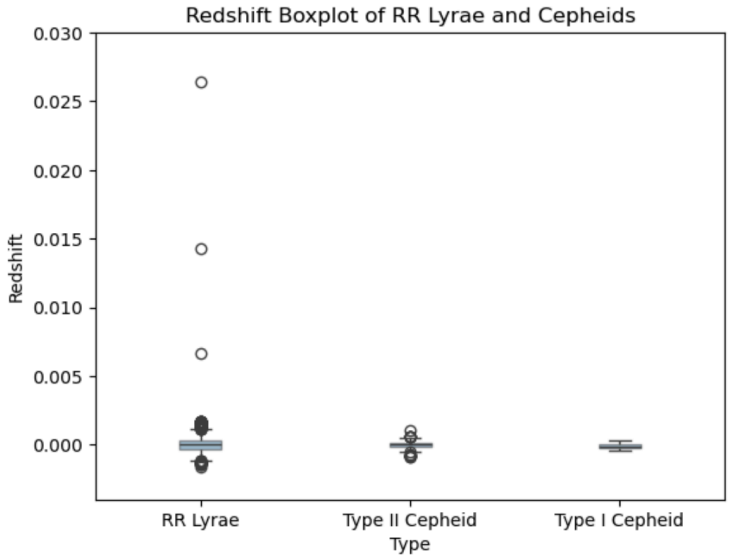

Fig 11: Box Plot Showing the Redshift Distributions for Each Type of Variable Star. This plot shows even more clearly the difference in range that makes this parameter so valuable.

A box plot was created to analyze the spread of the data points, better illustrating their variability. The range of the Type I Cepheid is minimal compared to the Type II Cepheid. These two, in turn, have a smaller range than the RR Lyrae variables. Although each of these three stars does not show a significant difference in terms of mean, this feature was included in the model solely due to the variations in ranges among the three groups.

We acknowledge that misclassification patterns were not fully analyzed in our study, but a confusion matrix was used to assess the model's performance. From this, we noticed that many errors occurred between Cepheid Type I and RR Lyrae stars. This is likely due to the overlap in redshift values, as shown in the strip plot above. In future research, we aim to better understand misclassified stars and examine them using outlier detection methods.

We also notice that several RR Lyrae stars have high redshift values. We see this in both plots, but we did not take any further steps to filter or address these outliers. These could be caused by various factors and are worth investigating to improve the model.

Machine Learning Classification

Because this classification task involves multiple parameters that cannot be used individually, a Random Forest Classifier and a Gradient Boosting Classifier were employed for the classification.

The Random Forest Classifier makes classifications by using multiple decision trees and arriving at a single result (IBM, 2023). Because it combines the predictions of several decision trees, it is less likely to overfit and more capable of analyzing and recognizing complex patterns or conditions within the data, making it a more suitable choice for this classification task. Gradient Boosting, on the other hand, is a classification model that combines multiple weak models to build a stronger, more accurate model (Scikit-learn, 2009). These weak models are often based on a single or a few parameters with weak correlations, and they can be combined to form a more robust model that captures various patterns. We did not include neural networks, as they require larger and more balanced datasets than the one used in our research. We encourage future research to explore neural networks, which might reveal additional insights into variable star classification. Additionally, we did not incorporate SHAP or LIME in this study due to scope limitations. However, both approaches should be considered in future research because of their numerous benefits.

Since no clear pattern was found using color indices or galactic longitude, these features were omitted. Instead, the other parameters—galactic latitude, redshift, and metallicity—were utilized. Any null values in the columns were replaced with the column median using SimpleImputer. The median was chosen because several outliers in each parameter could skew the mean, making it less suitable for representation. Therefore, the median was used to fill all missing values. We acknowledge that more advanced methods, such as MICE, could have improved the model's robustness. We intend to explore this in future research and model iterations. Future studies might also benefit from using GridSearchCV or Bayesian Optimization.

RESULTS

The Random Forest model was tuned using 100 n-estimators (or decision trees) to identify patterns in the dataset while minimizing the risk of overfitting. Generally, increasing the number of n-estimators adds more decision trees and can improve accuracy; however, too many n-estimators may lead to overfitting and reduce the model’s performance. The accuracy of this model was 92%. In comparison, the gradient boosting model achieved an accuracy of 89%. Additionally, we employed hyperparameter optimization to further enhance performance. Using GridSearchCV, the optimized model reached an accuracy of 91.5%.

The tuned Gradient Boosting model resulted in an accuracy of 90.5%. To provide context, we also developed a k-nearest neighbors (k-NN) model. Its accuracy was 89.7% with a deviation of 1.4%. Meanwhile, our random forest classifier achieved an accuracy of 92.4% with a deviation of 0.7%. These comparisons demonstrate the higher performance. of the model we programmed compared to the k-NN model.

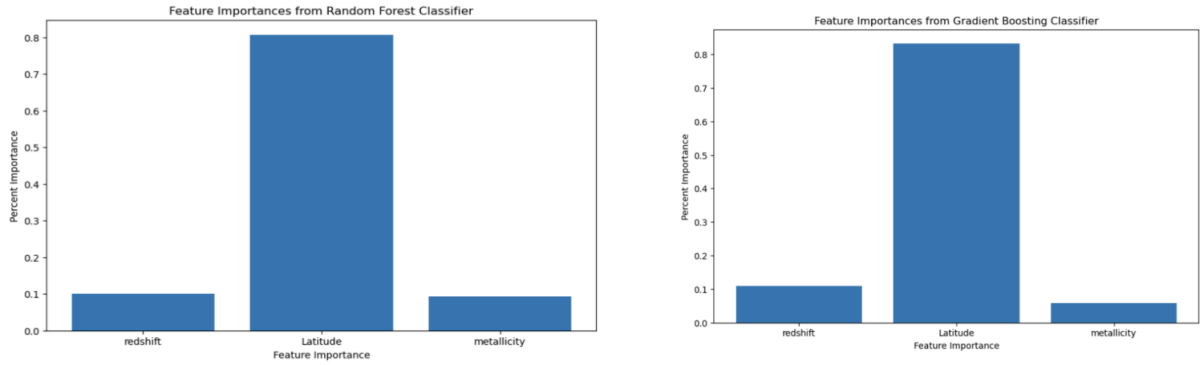

Fig 1: Feature Importance Graphs for Each Machine Learning Model. The graphs shown emphasize the importance of the 3 features chosen and especially the amount of weight held by galactic latitude for classification. However, there is a discrepancy between whether the model gave more importance to the redshift of metallicity, but the higher accuracy of the random forest model suggests that metallicity might be the more important parameter.

As shown in the graphs above, both models identified galactic latitude as the key factor for classification. While the Gradient Boosting classifier considered redshift more important than metallicity, the Random Forest model regarded metallicity as a more significant distinguishing feature than redshift.

We concluded that the Random Forest model had higher accuracy than the Gradient Boosting model because it considered metallicity more important than redshift, as metallicity showed clearer differences between star categories, unlike redshift. This conclusion was further supported by removing galactic latitude from both models and plotting feature importance graphs for each. The following results were obtained:

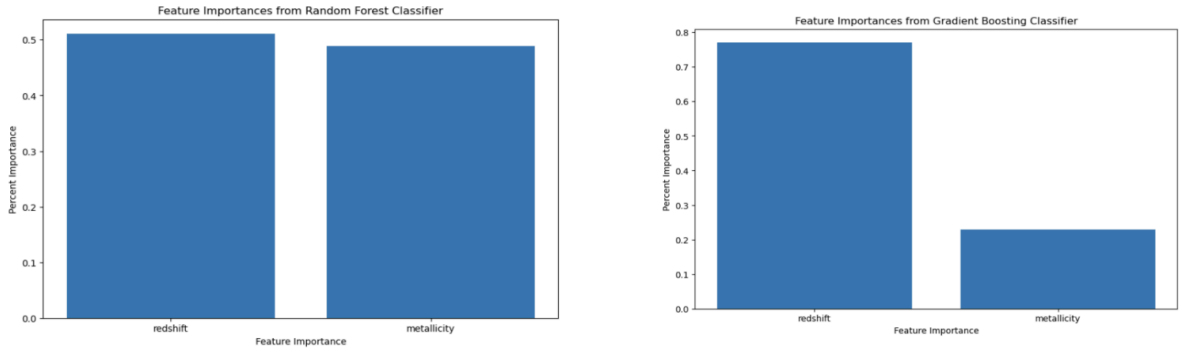

Fig 2. Feature Importance Graphs with Galactic Latitude Excluded. To see clearly the difference in weightage given to redshift and metallicity, galactic latitude was removed, and we see that under that condition, metallicity is given less weightage in both models.

As shown in the graph, the Random Forest model emphasizes metallicity more than the Gradient Boosting model. This contributed to its higher accuracy of 70%, compared to the Gradient Boosting

Model's accuracy of 67%. To justify our feature selection, we applied a statistical method using mutual information scores via the SelectKBest algorithm. This confirmed that galactic latitude, metallicity, and redshift are the most important factors in classification. Using these features, we trained the model and evaluated it with precision, recall, and F1-score. The model achieved an accuracy of 93%, with precision scores of 0.90 and 0.94, F1 scores of 0.87 and 0.95, and recall values of 0.84 and 0.94 for Cepheid and RR Lyrae stars, respectively. This indicates the model was better at avoiding false positives, as shown by the high precision scores, and false negatives, as shown by the high recall scores. The high F1 score suggests a good balance between precision and recall.

While we presented feature importance plots for both our classifiers, we acknowledge that a detailed analysis of these plots was not performed. Specifically, we did not use SHAP (SHapley Additive exPlanations) or permutation importance techniques to analyze how features interact, which we hope future research will explore. We also recognize that our models currently lack confidence intervals or uncertainty estimates. Future work should address this by integrating Bayesian models and Monte Carlo dropout to improve the model's confidence and accuracy. Additionally, our analysis focused on global importance, with galactic latitude, metallicity, and redshift emerging as major contributors. Using SHAP values to examine individual prediction contributions and potential feature interactions could uncover patterns that were previously missed.

DISCUSSIONS

Our research resulted in the development of two classification models with 92% and 89% accuracy using Random Forest and Gradient Boosting algorithms, respectively. These accuracy figures come from models that were not optimized and did not involve imputation. The study demonstrates meaningful relationships between star classification and properties such as redshift and metallicity, which support our initial hypothesis. However, our study has limitations that should be addressed in future research. The color indices obtained through the color filters we used showed no significant relationship that could help in classifying stars. Our findings suggest that features other than light curves are correlated with the type of variable star. This work could be valuable in the field of variable star classification, where these additional parameters, beyond light curves, might be incorporated into existing models (Bhardwaj et al., 2023) or used independently to improve the accuracy of variable star classifiers.

At the start of our study, we hypothesized that the following parameters — redshift, color indices, galactic coordinates, and metallicity — would all be important for the classification model. However, our findings suggest that galactic coordinates and color indices weren’t as helpful in classification.

With our initial model, we achieved a very low accuracy rate. To improve this, we added an extra classification group—splitting Cepheid Variables into Type I and Type II Cepheid Variables—which greatly increased the accuracy from 69% (our original rate) to 92%. The classification report we generated shows that our model has a high precision score of 0.99 and a solid F1-score of 0.81 for Type I Cepheids, indicating reliable classification for this group. However, the model performs less well when classifying RR Lyrae stars and Type II Cepheids. It received precision and recall scores of 0.00 for these classes, likely due to the small number of stars in those categories. These results highlight the effect of

class imbalance in our dataset and suggest that future work should consider resampling methods such as SMOTE.

After training our original model, we used multiple strategies to ensure robustness. We then identified outliers using the interquartile range method and removed them from the redshift parameter, which improved our model's stability. To estimate accuracy, we employed 10-fold cross-validation, resulting in the following: Random Forest accuracy of 91.5% with a 1.2% deviation; Gradient Boosting accuracy of 90.5% with a 1.4% deviation. These accuracy metrics come from models that handled and removed outliers in the data.

In our study, the dataset was randomly divided into an 80/20 training and testing set, and metrics were obtained using k-fold cross-validation, which helps reduce the risk of overfitting. However, we acknowledge the lack of external validation and recommend that future researchers explore this approach. Future studies should also incorporate data from other databases, such as Gaia or ASSA-SN, to enhance robustness. We also encourage future research to investigate other types of variable stars in classification.

CONCLUSION

The research that we've conducted leads us to the conclusion that there is a correlation between the parameters of galactic latitude, metallicity, and redshift and the type of variable star. The research has demonstrated the importance of these three features in machine learning classification of variable stars as well, which further validates the conclusion that there is a correlation between these parameters and the variable star type.

However, due to the limitations of this study, such as the lack of hyperparameter tuning and confidence intervals, our conclusion should be explored further in the future. The use of these methods and other techniques will enhance the robustness and generalizability of our study's findings.

ACKNOWLEDGEMENTS

We thank Dr. Shyamal Mitra at the University of Texas at Austin for his guidance and support throughout this paper. We also thank the peer mentors of the High School Research Academy at UT Austin—Anthony Yang, Juhi Malwade, and Hanyu Wei—for their help and advice during our project. Additionally, we appreciate Dr. Karl Gebhardt, the chairman of the Astronomy Department at UT Austin, for his constructive feedback.

REFERENCES

Bhardwaj, A., Bellinger, E. P., Kanbur, Shashi M, & Marconi, M. (2023). *Predicting Physical Parameters of Cepheid and RR Lyrae variables in an Instant with Machine Learning*. ArXiv.org. https://arxiv.org/abs/2303.13692

Color Indices and Surface Temperature. (n.d.). Www.pas.rochester.edu. https://www.pas.rochester.edu/~blackman/ast104/cindex.html

IBM. (2023). *What Is Random Forest? | IBM*. Www.ibm.com; IBM. https://www.ibm.com/topics/random-forest

Kim, D.-W., & Coryn A. L. Bailer-Jones. (2015). A package for the automated classification of periodic variable stars. *Astronomy and Astrophysics*, *587*, A18–A18. https://doi.org/10.1051/0004-6361/201527188

Kuntzer, T., Tewes, M., & Courbin, F. (2016). Stellar classification from single-band imaging using machine learning. *Astronomy & Astrophysics*, *591*, A54. https://doi.org/10.1051/0004-6361/201628660

Las Cumbres Observatory. (2023). *Redshift | Las Cumbres Observatory*. Lco.global. https://lco.global/spacebook/light/redshift/

Metallicity of stars. (n.d.). Icc.dur.ac.uk. https://icc.dur.ac.uk/~tt/Lectures/Galaxies/TeX/lec/node27.html

NASA. (2018). *The Cosmic Distance Scale*. Nasa.gov. https://imagine.gsfc.nasa.gov/features/cosmic/milkyway_info.html

Scikit-learn. (2009). *3.2.4.3.5. sklearn.ensemble.GradientBoostingClassifier — scikit-learn 0.20.3 documentation*. Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

SIMBAD Astronomical Database - CDS (Strasbourg). (n.d.). Simbad.cds.unistra.fr. https://simbad.cds.unistra.fr/simbad/

Sky Maps with Pierre Auger Data. (2025). Auger.org. https://www.auger.org/education/Auger_Education/galacticcoordinates