An Analysis of Machine Learning-Based Classification of Variable Stars

Anjie Liu[1] and Jasmine C. Xu[1]

[1] The University of Texas at Austin

1. RESEARCH QUESTION

Our research aims to determine the applicability of machine learning models in the effective classification of variable stars. Variable stars are stars who exhibit intermittent fluctuations in luminosity over time. These fluctuations can be attributed to various sources, resulting in distinct patterns of period and amplitude variations. As such, it is reasonable to use features relating to period and magnitude as parameters for a machine learning model that aims to differentiate variable stars into their respective classes.

The relevance of this question comes with the increasing pace and volume of astronomical data collection by state-of-the-art telescopes that have rendered the conventional manual classification approach impractical. Hence, it is sensible to address astronomical questions using modern techniques that can handle vast amounts of data with reasonable efficiency and accuracy. Additionally, accuracy can be maximized by comparing the efficacy of multiple machine learning models; and robustness can be achieved by applying the models to different variable star databases.

2. BREAKING UP THE RESEARCH QUESTION

Our research question can best be approached by splitting it into sub-questions, and addressing these before we answer the question as a whole.

2.1. *Obtaining a source of data*

One of our first questions was to determine a database of variable stars to train our models. Upon consideration of our options, as well as of past research, we decided to utilize data from the fourth and most recent phase of the Optical Gravitational Lensing Experiment (OGLE-IV). OGLE uses microlensing to identify stars in the Small and Large Magellanic Clouds, as well as in the Galactic Bulge and Galactic Disk, providing one of the most comprehensive and up-to-date databases of classified variable stars (Udalski et al. 1992). Additionally, OGLE-IV includes an extensive collection of features for each star, including mean I-band and V-band magnitudes, I-band amplitudes, and photometric data. This wealth of features makes the database well-suited as a source of input data for supervised classifier models, and the database would be well-suited for our purposes.

Ultimately, we extracted a dataset comprised of around 736,000 variable stars taken from all target fields of the OGLE-IV database. This encompasses eight types of variable stars, namely, eclipsing binaries, RR Lyrae variables, long-period variables, Delta Scuti variables, classical Cepheids, type II Cepheids, Heartbeat stars, and anamolous Cepheids. The original distribution of these star types in our dataset is shown in **Table 1**.

Variable Star Type	Count	Percent of Total
Eclipsing Binaries	499,203	67.83%
RR Lyrae Variables	128,273	17.43%
Long-Period Variables	65,981	8.97%
Delta Scuti Variables	27,392	3.72%
Classical Cepheids	11,703	1.59%
Type II Cepheids	2,010	0.27%
Heartbeat Stars	991	0.13%
Anamolous Cepheids	389	0.053%
	735,942	100%

Table 1. Counts and percentages of each variable star type in data set

It can be observed from **Table 1** that there is a large imbalance of classes within our data set. Eclipsing binaries represent over half of the entire data set, while all 7 other classes represent less than 20%. Notably, Type II Cepheids, Heartbeat stars, and anomalous Cepheids have fewer than 10,000 observations. This is better visualized in **Figure 1**. This observation leads to the question of how this imbalance should be managed in the training of our classifiers; this question is to be addressed in subsection 2.4.

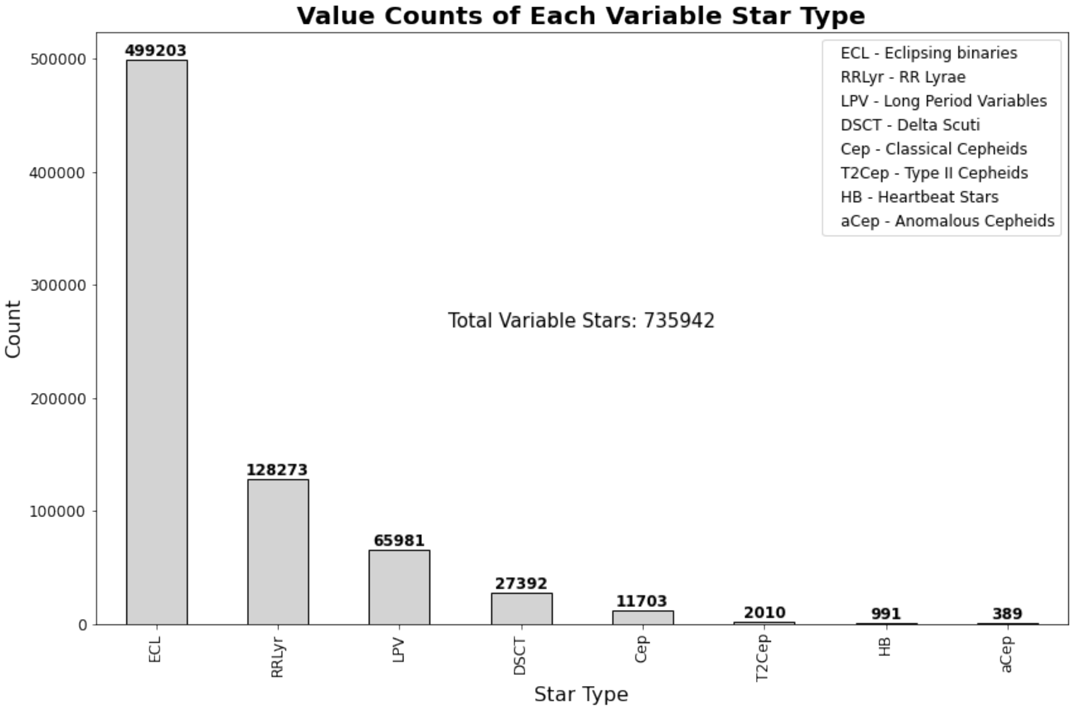


Figure 1. Barplot of star counts, demonstrating class imbalance

2.2. *Determining appropriate machine learning algorithms*

Our next sub-question was to determine the most appropriate machine learning models for our task. Given that the task at hand involves the classification of variable stars, it naturally requires the construction of classifier models who group input data into a discrete set of categories. Such models include random forests, support vector machines, k-nearest neighbors, and so on. We ultimately chose to build a random forest classifier and an XGBoost classifier, both of which are ensemble learning methods who combine the predictions of multiple learners to obtain high prediction accuracies (Breiman 2001).

The robustness of these algorithms, most especially the random forest classifier, make them a reasonable starting point for our research. Future steps in our research will likely involve the construction of more machine learning models that will be compared to the current two. For example, in addition to mean magnitudes and amplitudes, OGLE-IV also provides photometric data for each variable star, which would be applicable as input data for a recurrent neural network.

2.3. *Selecting relevant features*

With the abundance of features given to us by OGLE-IV, it next became imperative to choose the optimal features to use as predictors of variable star type.

The first step taken in this process was to eliminate features containing a high proportion of missing data to prevent the loss of valuable training data. All features with over 20% of its data missing were dropped from the dataset. Our original dataset of close to 736,000 rows and 29 numerical features was cut down to around 670,000 rows and 6 features following this procedure.

A majority of the eliminated rows can be attributed to the exclusion of long-period variables, due to the fact that the T1_0 feature - the time of minimum brightness - is unfilled for each star. Consequently, all 65,981 long-period variables were removed from the dataset. Because we deem the time of minimum brightness to be an important determining feature of the type of variable star, we will not aim to classify long-period variables in this investigation.

The amount of features eliminated from the dataset is due to a similar reason - many features are specific to a certain type of variable star, and while they are filled for all occurrences of that variable star type, it is missing for all other types.

The remaining six features included four that described amplitude, magnitude, and period, while the other two features represented right ascension and declination. Based on our knowledge of the field, we deemed the two features that described equatorial coordinates to be insignificant to the model classification of variable star type, and so they were removed. We decided to make use of all four remaining features.

With these chosen four features, a correlation matrix was constructed to visualize the relationships between each feature, and their ability to predict the target variable. For this purpose, the variable star type - our target variable - was encoded as numerical values using One Hot Encoding. One Hot Encoding encodes each unique value within the target variable as a new column, to avoid ordinality within the target variable.

Our selected features and their definitions are shown in Table 2.

Feature	**Description**
T0_1	time of minimum brightness, in days
A_1	I-band amplitude, or main eclipse depth
I	mean (for pulsating stars) or maximum (for eclipsing binaries) I-band magnitude
P_1	period, in days

Table 2. Descriptions of features for use in the training of machine learning classifier models

2.4. *Addressing class imbalance*

As previously mentioned, there is a significant amount of class imbalance present within our dataset that could greatly hinder a classifier model's ability to correctly identify minority classes. We aim to address this class imbalance using Synthetic Minority Oversampling Technique (SMOTE), a type of data augmentation that generates synthetic samples for the minority classes using the existing minority class samples (Chawla et al. 2011). This method will ensure that there are equal numbers of stars in each class, and ultimately increase the model's ability to learn from the minority classes.

The distribution of star types in the resampled dataset are shown in Table 3, and can be compared to the largly class-imbalanced dataset as shown in Table 1. Notice also that, as previously mentioned, long-period variables have been excluded from the investigation.

3. BUILDING MACHINE LEARNING MODELS

The final step in the preprocessing of our data was to encode the target variable - the type of variable star - in a way such that their labels can be predicted. We used Scikit-learn's LabelEncoder to assign a unique numerical value

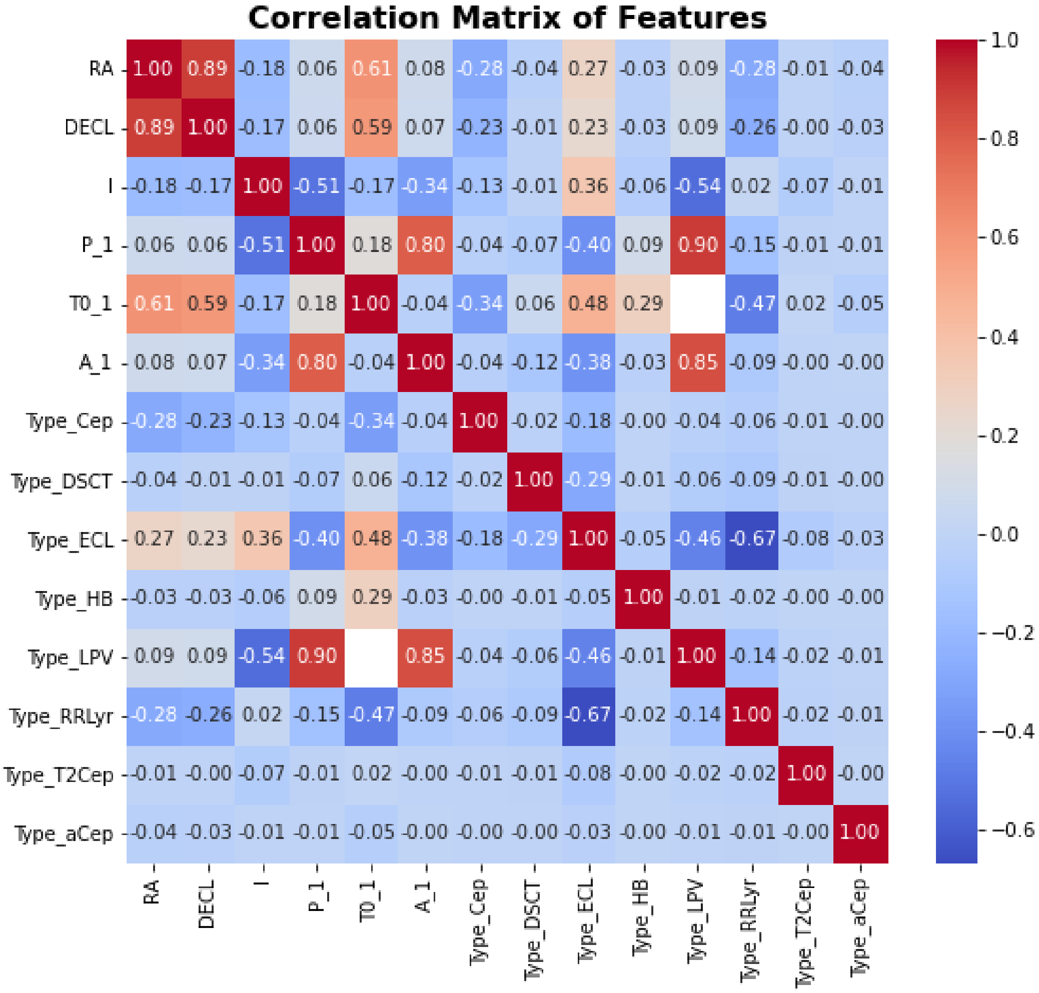


Figure 2. Correlation matrix of potentially relevant features

Variable Star Type	Count	Percent of Total
Eclipsing Binaries	499,181	14.29%
RR Lyrae Variables	499,181	14.29%
Delta Scuti Variables	499,181	14.29%
Classical Cepheids	499,181	14.29%
Type II Cepheids	499,181	14.29%
Heartbeat Stars	499,181	14.29%
Anamolous Cepheids	499,181	14.29%
	3,496,388	100%

Table 3. Counts and percentages of each variable star type in the resampled data set

to each variable star type. The limitation to this method, however, is that the classifier model may perceive these labels to be inherently ordinal. An alternative method to LabelEncoder is One-Hot Encoding, which avoids the issue of assumed ordinality by creating a new column in the dataset for each unique label. However, we decided against this method because it would greatly increase the dimensionality of our data.

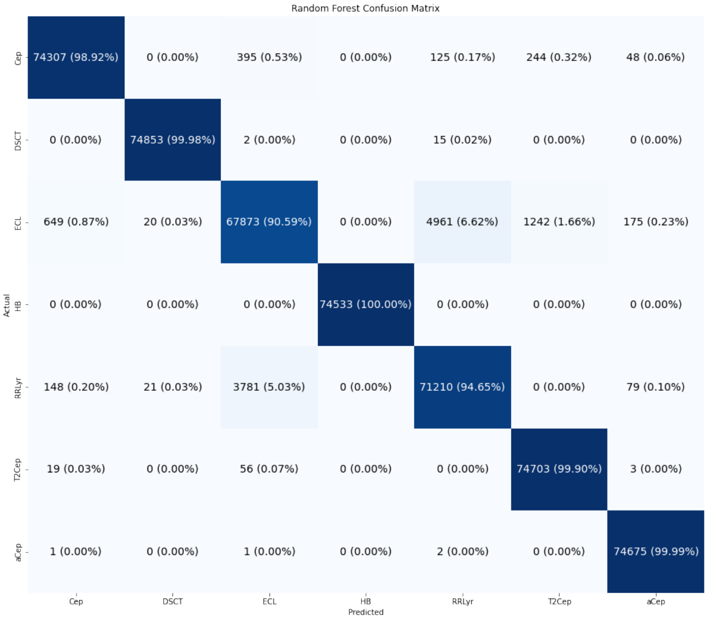
Following the preprocessing of our dataset, we constructed a random forest model and an XGBoost model. The construction, training, and testing of our machine learning models relied heavily on Python's Scikit-learn library.

This first required us to split our data into training, testing, and validation sets. The training set consisted of 70% of the dataset, while the testing and validation sets each contain 15% of the data. The validation set will be used to determine hyperparameters for the classifiers. As a result of the splitting, 2,445,986, 524,140, and 524,141 stars (genuine and synthetically-generated) were assigned to the training, testing, and validation sets, respectively.

To optimize model performance, a randomized search was used to tune hyperparameters for both models. The determined optimal parameters were then used to train the respective models.

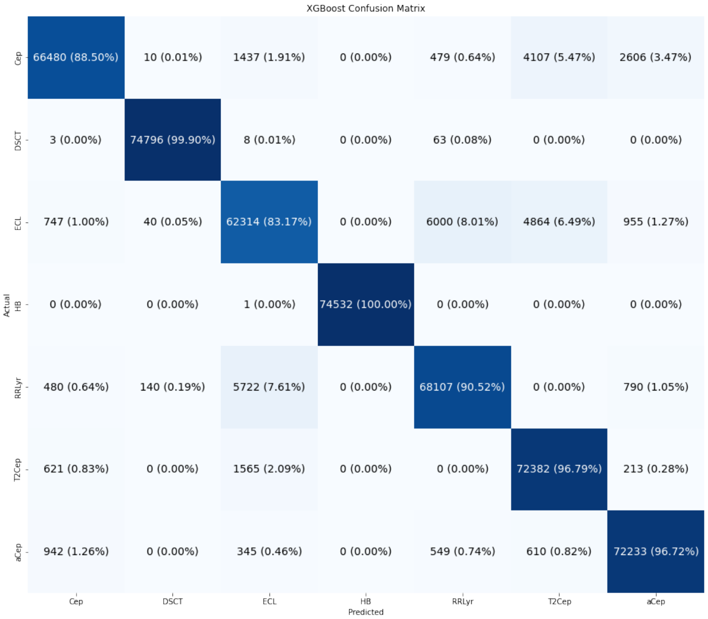
4. RESULTS

Our random forest and XGBoost classifiers achived accuracy scores of 97.71% and 93.65%, respectively. Their confusion matrices, shown in **Figure 3** and **Figure 4**, display comparable overall trends.



	Cep	DSCT	ECL	HB	RRLyr	T2Cep	aCep
Cep	74307 (98.92%)	0 (0.00%)	395 (0.53%)	0 (0.00%)	125 (0.17%)	244 (0.32%)	48 (0.06%)
DSCT	0 (0.00%)	74853 (99.98%)	2 (0.00%)	0 (0.00%)	15 (0.02%)	0 (0.00%)	0 (0.00%)
ECL	649 (0.87%)	20 (0.03%)	67873 (90.59%)	0 (0.00%)	4961 (6.62%)	1242 (1.66%)	175 (0.23%)
HB	0 (0.00%)	0 (0.00%)	0 (0.00%)	74533 (100.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
RRLyr	148 (0.20%)	21 (0.03%)	3781 (5.03%)	0 (0.00%)	71210 (94.65%)	0 (0.00%)	79 (0.10%)
T2Cep	19 (0.03%)	0 (0.00%)	56 (0.07%)	0 (0.00%)	0 (0.00%)	74703 (99.90%)	3 (0.00%)
aCep	1 (0.00%)	0 (0.00%)	1 (0.00%)	0 (0.00%)	2 (0.00%)	0 (0.00%)	74675 (99.99%)

Figure 3. Confusion matrix of random forest classifier



	Cep	DSCT	ECL	HB	RRLyr	T2Cep	aCep
Cep	66480 (88.50%)	10 (0.01%)	1437 (1.91%)	0 (0.00%)	479 (0.64%)	4107 (5.47%)	2606 (3.47%)
DSCT	3 (0.00%)	74796 (99.90%)	8 (0.01%)	0 (0.00%)	63 (0.08%)	0 (0.00%)	0 (0.00%)
ECL	747 (1.00%)	40 (0.05%)	62314 (83.17%)	0 (0.00%)	6000 (8.01%)	4864 (6.49%)	955 (1.27%)
HB	0 (0.00%)	0 (0.00%)	1 (0.00%)	74532 (100.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
RRLyr	480 (0.64%)	140 (0.19%)	5722 (7.61%)	0 (0.00%)	68107 (90.52%)	0 (0.00%)	790 (1.05%)
T2Cep	621 (0.83%)	0 (0.00%)	1565 (2.09%)	0 (0.00%)	0 (0.00%)	72382 (96.79%)	213 (0.28%)
aCep	942 (1.26%)	0 (0.00%)	345 (0.46%)	0 (0.00%)	549 (0.74%)	610 (0.82%)	72233 (96.72%)

Figure 4. Confusion matrix of XGBoost classifier

Accuracy scores for both models are high, which suggests that despite the low number of features, the star types are different enough for their variations to reflect in these few features. For example, the summary statistics of the P_1 variable, shown in Table 4, show visible distinctions between the periods of star types.

Type	Mean	Standard deviation	Minimum	Median	Maximum
Cep	3.40	4.89	0.22	2.24	208.80
DSCT	0.09	0.03	0.03	0.08	0.30
ECL	6.17	35.37	0.05	0.67	4,200.00
HB	291.34	183.84	2.79	265.32	2,555.50
RRLyr	0.50	0.13	0.20	0.52	1.00
T2Cep	11.81	12.03	0.79	8.37	84.81
aCep	1.05	0.45	0.37	0.94	2.67

Table 4. Summary statistics for the P_1 variable, showing star period in days

It also appears that accuracy scores are somewhat negatively correlated with the rarity of variable star types in the original imbalanced dataset. In other words, star types originally present in higher proportions have lower accuracy scores for both classifer models, and vice versa. For example, heartbeat stars made up just 0.053% of the original dataset, and were classified completely correctly by both models. On the other hand, eclipsing binaries comprised a majority of the original dataset, and were classified by the random forest and XGBoost models with accuracies of 90.59% and 83.17%, respectively.

5. RELEVANT WORK

Machine learning-based classification of variable stars is not a novel idea. Prior research has been conducted in this area using different machine learning algorithms, different variable star databases, and so on. It is reasonable to compare our findings to that of prior research, such that areas of weakness can be addressed.

Hosenie et al. (2019) has done similar work regarding applying machine learning models to variable star classification but with a different set of data: Catalina Real-Time Transient Survey (CRTS). They implemented three machine learning algorithms, namely random forest, decision tree, and k-nearest neighbors. Although the random forest model achieves the best performance, all three models perform poorly overall due to the lack of data for certain classes such as delta Scuti and ACEP. Therefore, they propose a mechanism of decomposing a multiclass problem into several binary classification steps for improved results. In addition, they develop a hierarchical approach to further group the data by similarities, even though the success is not consistent for all 11 classes they are investigating.

While the previous work provides insights on building an automated pipeline of classifying variable stars, Zorich et al. (2020) further advance the field with more solidary work. They tackle the expensive and time-consuming offline machine learning model traning issue that arises due to a streaming wealth of data provided by advanced telescopes. Therefore, in this work, they propose a streaming probabilistic classification model that updates itself in real time with new observations.

6. REFLECTION

Our random forest and XGBoost models achieve high accuracy scores of 97.71% and 93.65%, respectively. These satisfactory results across all variable star types are largely aided by the choices of features, as well as the effective management of imbalanced data.

While our investigation as of now has shown positive results, we hope to extend it by constructing more machine learning models that can be compared to our current random forest and XGBoost classifiers. We most strongly aim to take advantage of the availability of photometric data to implement a recurrent neural network. Neural networks, in contrast to our current models, are not ensemble methods, and so would provide a new approach to variable star classification. Similarly, a convolutional neural network could be applied alongside this, given that photometric data can be represented as images.

Another further step we aim to take is to apply our current models to other variable star databases in addition to OGLE-IV, which our models were trained upon. If applied to classified datasets, this would allow us to confirm the applicability and effectiveness of our models; and if applied to unclassified datasets, this gives us the potential of identifying new variable stars.

REFERENCES

Breiman, L. 2001, Machine Learning, 45, 5, doi: 10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2011, doi: 10.48550/ARXIV.1106.1813

Hosenie, Z., Lyon, R. J., Stappers, B. W., & Mootoovaloo, A. 2019, Monthly Notices of the Royal Astronomical Society, 488, 4858, doi: 10.1093/mnras/stz1999

Udalski, A., Szymanski, M., Kaluzny, J., Kubiak, M., & Mateo, M. 1992, Acta Astronomica, 42, 253. https://ui.adsabs.harvard.edu/abs/1992AcA....42..253U

Zorich, L., Pichara, K., & Protopapas, P. 2020, Monthly Notices of the Royal Astronomical Society, 492, 2897, doi: 10.1093/mnras/stz3426