Draft version October 12, 2023
Typeset using LaTeX default style in AASTeX631

Machine Learning-Based Classification of Variable Stars: Data Collection

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 celestial objects characterized by their 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 classification of astronomical objects is a sensible decision in the modern context, where the pace and volume of astronomical data collection from state-of-the-art telescopes such as the James Webb Space Telescope have rendered the conventional human-driven classification approach as increasingly impractical. Hence, it makes sense to address astronomical questions using modern techniques that can handle cast amounts of data with reasonable efficiency and accuracy.

2. SOURCES OF DATA

To address our research question, we have obtained 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.

3. DATA STATISTICS AND ANALYSIS

Our data set is comprised of over 700,000 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.

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%

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 the following barplot.

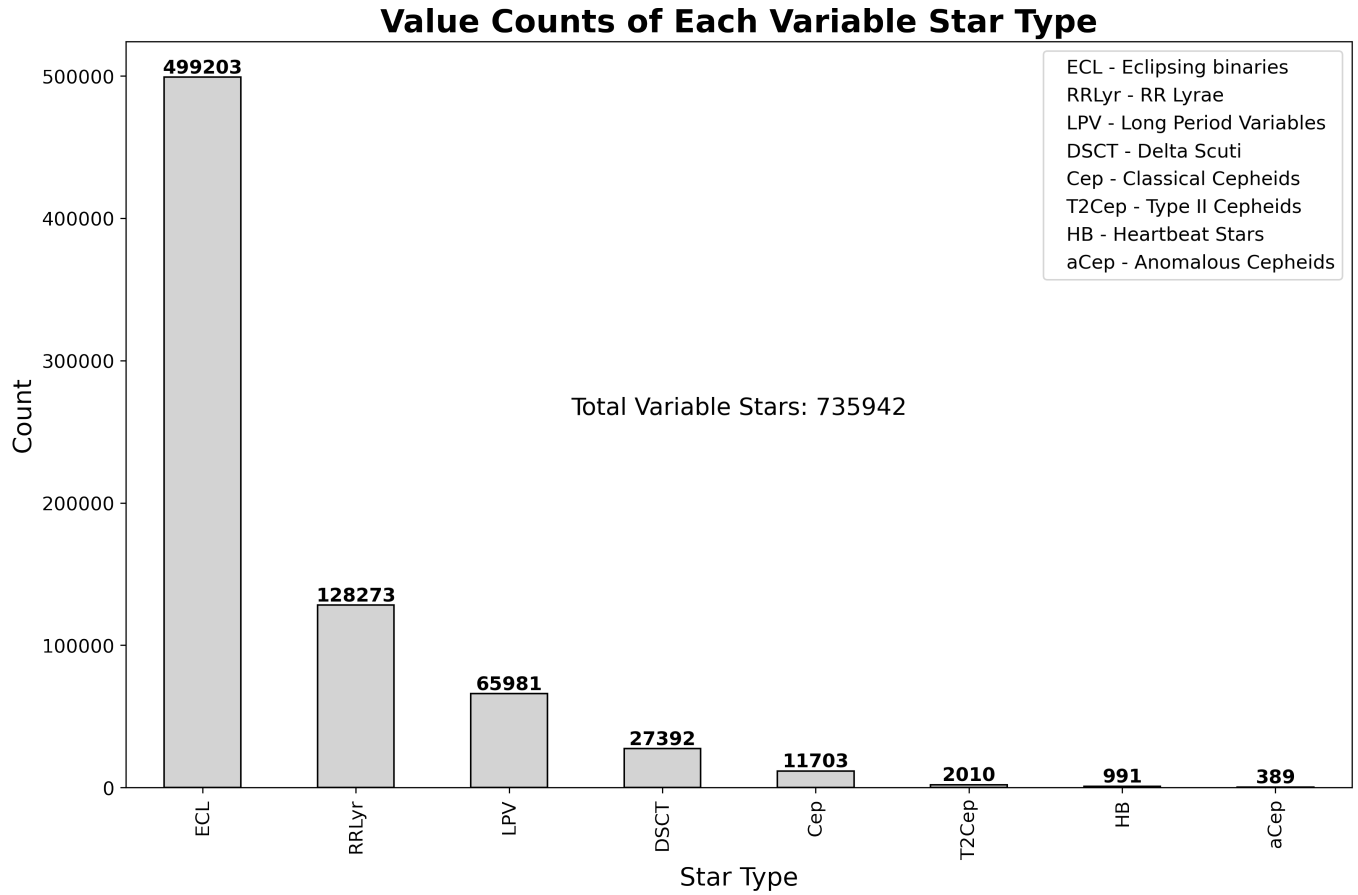

Figure 1. Barplot of star counts, demonstrating class imbalance

This class imbalance can have various consequences on our results. The resulting model may become biased, such that it performs better when classifying the majority class, and performs poorly on the minority classes. The model may yield a high accuracy score despite consistent mis-classifications of the minority class, because the score is skewed substantially by the correct classifications of the majority class. The model may exhibit low accuracy scores when given unseen data that contains members of the minority class. We will further address this limitation in a later section.

Based on relevance and availability, we have chosen a subset of four features for use in our machine learning classifiers. Features with over 20% of their total length encoded as missing (NaN) values were eliminated from contention as predictor features, in order to prevent the loss of too much valuable training data. The original 29 numerical features in the dataset was cut down to just six after these NaN-populated features were filtered out. These 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.

With the remaining 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. This is the same method that will be used to encode the variable star type during the training process of the classifier models. One Hot Encoding encodes each unique value within the target variable as a new column, to avoid ordinality within the target variable.

The feature correlation matrix is displayed below.

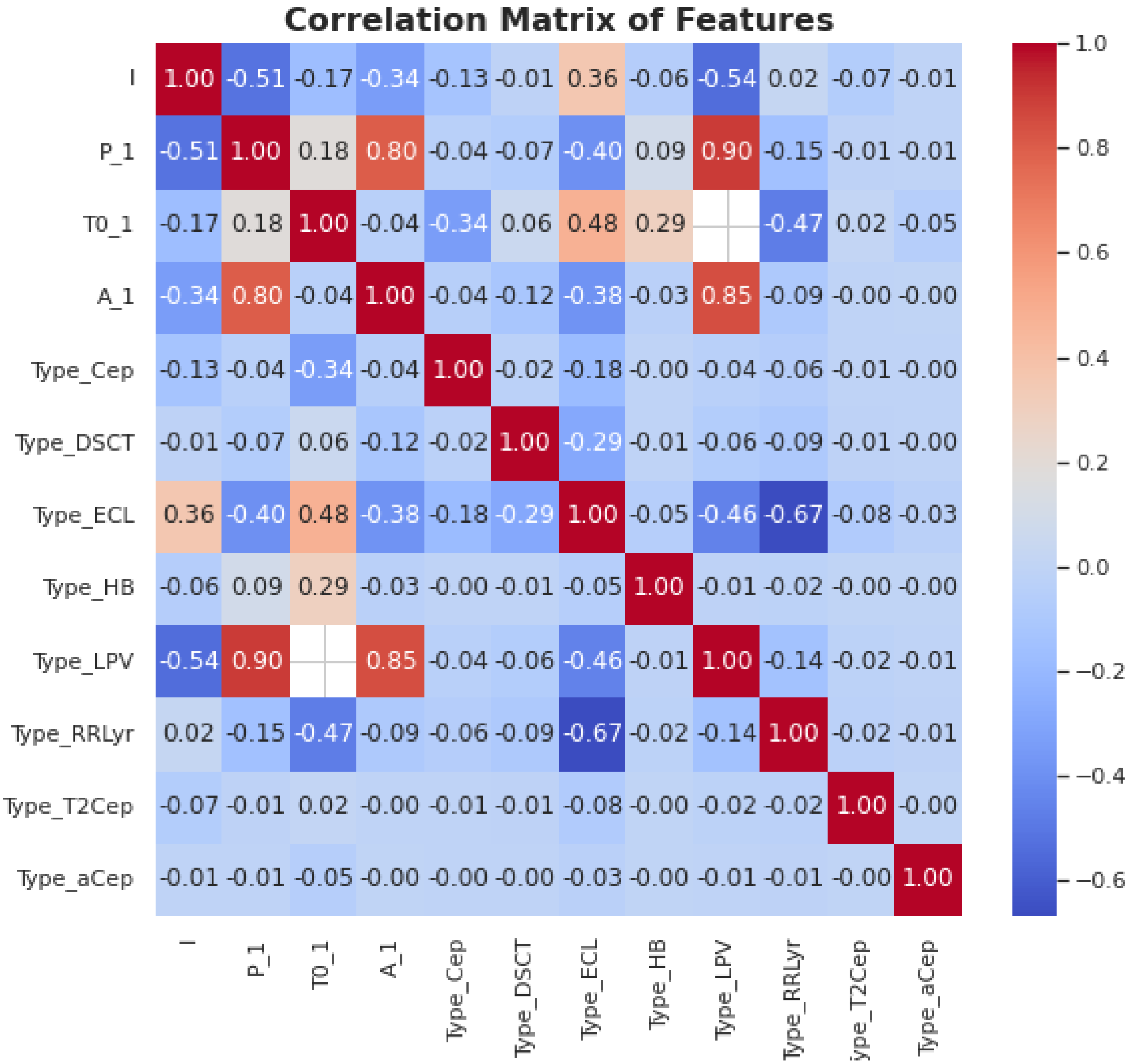

Figure 2. Correlation matrix of relevant features, showing relationship between each feature and the target variable

Our selected features and their definitions are as follows:

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

Below are the summary statistics for each feature, split by variable star type. Substantial variation can be seen from these statistics, and as such, a machine learning model designed for classification is well-suited to distinguish and categorize the various types of variable stars accurately. Such a model can leverage these statistical differences in feature distributions to make informed predictions, ultimately aiding in the efficient and accurate classification of variable stars.

T0_1

	mean	std	min	50%	max
Type					
Cep	6,178.00	381.65	6,000.00	6,001.19	7,061.61
DSCT	7,000.04	0.03	7,000.00	7,000.04	7,000.23
ECL	7,003.09	20.68	7,000.00	7,000.34	9,696.00
HB	9,143.62	137.82	9,000.05	9,112.20	10,621.90
RRLyr	6,638.20	480.59	6,000.00	7,000.10	7,000.94
T2Cep	7,005.85	7.64	7,000.00	7,002.48	7,048.55
aCep	6,329.56	470.46	6,000.00	6,000.73	7,001.94

A_1

	mean	std	min	50%	max
Type					
Cep	0.37	0.19	0.01	0.34	1.38
DSCT	0.20	0.11	0.01	0.18	0.99
ECL	0.43	0.29	0.00	0.37	4.91
HB	0.05	0.05	0.01	0.03	0.66
RRLyr	0.47	0.21	0.01	0.45	1.71
T2Cep	0.55	0.27	0.04	0.55	3.13
aCep	0.54	0.20	0.11	0.50	0.98

I

	mean	std	min	50%	max
Type					
Cep	15.66	1.33	10.23	15.69	19.83
DSCT	17.22	1.76	10.62	17.27	21.46
ECL	17.77	1.25	11.53	18.00	21.18
HB	14.82	0.99	12.57	14.74	18.19
RRLyr	17.44	1.44	10.79	17.63	21.27
T2Cep	15.22	1.79	10.63	15.30	19.75
aCep	16.84	1.34	12.36	17.29	18.94

P_1

	mean	std	min	50%	max
Type					
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 3. Summary statistics for the T0_1, A_1, I, and P_1 features

The distribution of features as shown in Table 3 indicate that each feature has vastly different ranges. Hence, it would be beneficial to normalize each feature between 0.0 and 1.0, such that training is less biased, and so that classifier models such as Random Forests are able to converge more quickly.

4. PRELIMINARY DATA ANALYSIS

An important preprocessing step in the training of machine learning models is the handling of missing data. We have elected to drop all observations in our dataset that contains one or more missing values. Our original dataset of close to 736,000 rows was cut down to around 670,000 rows 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.

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.

In the context of a Random Forest model, this issue could also be addressed by assigning class weights to each class, such that classes are weighed inversely proportional to their frequency in the dataset. In this way, the model is penalized more heavily for misclassification of rarer classes.

5. TOOLS USED IN DATA ANALYSIS

Our data analysis was conducted using several popular Python packages. The Pandas library was used to store our data in a dataframe, to extract relevant features and drop the remaining features, to remove rows with missing data, and also to compute summary statistics. The Matplotlib and Seaborn libraries were used to visualize our dataset.

6. DISCUSSION

Our preliminary data analysis has given us a wealth of insight into the variable star data from OGLE-IV. The improved dataset contains the most relevant information regarding the variable stars as all the selected parameters are only weakly correlated as shown in **Figure 2**.

The characteristics of each variable star type confirms with the existing knowledge. For example, heartbeat variable stars are known to have very small amplitude of less than 1 mmag; this characteristic is revealed in **Figure 4** when comparing the amplitude of each variable star type (Jayasinghe et al. 2021).

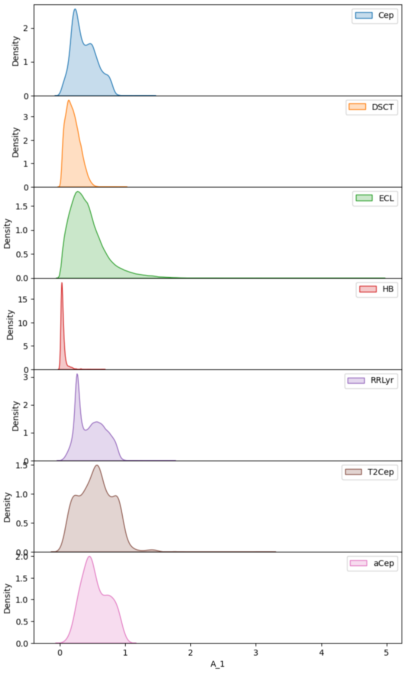

Figure 3. Density plot of the amplitude parameter of each variable star type

Overall, the cleaned OGLE-IV data prevents the issue of overfitting later in the machine learning models by handing missing values thoughtfully, adding weights on the low population group, and normalizing parameters.

Our next steps are to construct machine learning models using a subset of our dataset as training data. We aim to create a Random Forest and an XGBoost model, and to compare their accuracies. To add further complexity to our research question, we have also considered building a recurrent neural network that makes use of photometric data provided by the OGLE-IV database.

REFERENCES

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

Jayasinghe, T., Kochanek, C. S., Strader, J., et al. 2021, Monthly Notices of the Royal Astronomical Society, 506, 4083, doi: 10.1093/mnras/stab1920

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