








A Classification Catalog of Periodic Variable Stars for LAMOST DR9 Based on Machine Learning

Peiyun Qiao (乔佩云)^{1,2} , Tingting Xu (许婷婷)³ , Feng Wang (王锋)^{1,2} , Ying Mei (梅盈)^{1,2} , Hui Deng (邓辉)^{1,2} ,
Lei Tan (谈磊)^{1,2} , and Chao Liu (刘超)^{4,5} 

¹ Center for Astrophysics and Great Bay Center of National Astronomical Data Center, Guangzhou University, Guangzhou 510006, People's Republic of China;

fengwang@gzhu.edu.cn, meiying@gzhu.edu.cn, denghui@gzhu.edu.cn

² Peng Cheng Laboratory, Shenzhen, 518000, People's Republic of China

³ School of Mathematics and Computer Science, Yunnan Minzu University, Kunming, Yunnan, 650504, People's Republic of China

⁴ University of Chinese Academy of Sciences, Beijing, 100049, People's Republic of China

⁵ Key Laboratory of Space Astronomy and Technology, National Astronomical Observatories, Chinese Academy of Sciences, Beijing, 100101, People's Republic of China

Received 2023 October 3; revised 2024 March 3; accepted 2024 March 11; published 2024 April 16

Reporter: Lin Zhang (张琳)

Jun 21, 2024

Outline

- Introduction
- Data Preparation
- Classification Models
- Periodic Variable Source Identification and Classification in LAMOST DR9
- Conclusion

Introduction

- Since LAMOST entered routine observations, more than 19.46 million spectral data have been released. The LAMOST DR9 data set v1.06 was released in 2023 February. This data release comprises 11.21 million low-resolution spectra and 8.25 million medium-resolution spectra.
- However, the current LAMOST catalog lacks variable star information, which constrains the use of LAMOST data for studying variable stars.
- Several studies have been conducted using the LAMOST catalog to investigate variable stars. In some of these studies, variable sources are initially identified and crossmatched with published catalogs to determine their specific variable source types (e.g. [Tian et al. 2020](#); [Xu et al. 2022](#)).
- Although the types of variable sources obtained by the crossmatching method are credible, due to the limitation of the star catalog, many identified variable sources have not been classified accurately.

Introduction

- Several studies have also focused on identifying a particular type of variable source types (e.g. [Zhang et al. 2023](#); [Jia et al. 2023](#)). Identifying specific classes can satisfy scientists' needs, but much variable source information remains untapped in LAMOST.
- In recent years, machine-learning methods have been applied to classify variable sources and have shown significant advantages (Belokurov et al. 2003; Dubath et al. 2011; Jayasinghe et al. 2018; Narayan et al. 2018; Hosenie et al. 2019 Mahabal et al. 2017).
- Overall, the classification of variable sources for LAMOST still needs to be explored. In this study, we use machine learning to identify and classify variable sources in LAMOST DR9.

Data Preparation

- **2.1. Classification Feature Selection**

- Determining suitable variable star classification features is the key to further exploiting machine learning. In the study, we selected 30 features from two aspects.
- (1) Referring to previous work (Richards et al. 2011; Kim & Bailer-Jones 2016; Coughlin et al. 2021; van Roestel et al. 2021; Xu et al. 2022), we **selected 20 features** for variable source identification. These features are intrinsic statistical properties related to variable stars' scale, morphology, period, and other properties. They are calculated from a light curve's three vectors such as time, magnitude, and magnitude error.

Data Preparation

Table 1
Classification Feature Definitions

| Index | Features | Description |
|-------|--------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|
| 1 | Period | Best period. Lomb (1976), Scargle (1982), Kovács et al. (2002), Schwarzenberg-Czerny (1996), Stellingwerf (1978) |
| 2 | Kurtosis | Kurtosis of the magnitude distribution |
| 3 | Gskew | Median of magnitudes based measure of the skew |
| 4 | Nobs | Number of epochs in the light curve |
| 5 | Amplitude | Half of the difference between the median of the maximum and minimum 5% magnitudes |
| 6 | Shapiro Wilk | Shapiro–Wilk normality test statistics. Shapiro & Wilk (1965) |
| 7 | Median BRP | Fraction (≤ 1) of photometric points within amplitude/10 of the median magnitude. Richards et al. (2011) |
| 8 | Small Kurtosis | Small sample kurtosis of the magnitudes |
| 9 | ChiSquare | χ^2 value |
| 10 | Cusum | The range of magnitude cumulative sums. Kim et al. (2011) |
| 11 | Iter std | The data other than the median plus or minus twice the standard deviation are removed, and the standard deviation is recalculated. Xu et al. (2022) |
| 12 | MAD | Median absolute deviation |
| 13 | Eta | Von Neumann stat. Kim et al. (2014) |
| 14 | Stetson K | The Stetson K statistic. Stetson (1994) |
| 15 | Half mag amplitude ratio | HMAR. Ratio of magnitudes fainter or brighter than the average. Kim & Bailer-Jones (2016) |
| 16 | Beyond 1std ratio | Beyond 1. The fraction of observations beyond one standard deviation. D’Isanto et al. (2016) |
| 17 | Percent amplitude | PerAmp. Largest percentage difference between either the maximum or minimum magnitude and the median. Richards et al. (2011) |
| 18 | Std | Magnitude standard deviation |
| 19 | Skewness | The skewness of light curves |
| 20 | f1 power | $\frac{\chi_0^2 - \chi^2}{\chi_0^2}$ of the fit. van Roestel et al. (2021) |

Data Preparation

- (2) **Ten additional features** were selected from Gaia (Gaia Collaboration et al. 2021), ALLWISE (Wright et al. 2010), and the Two Micron All Sky Survey (2MASS; Skrutskie et al. 2006). High-precision photometric information is essential for variable star studies. We corrected for the extinction and reddening of **using the three-dimensional Bayestar19 reddening map (Green et al. 2019)**.

| | | |
|----|-----------|------------------------------------------------------|
| 21 | Parallax | Gaia parallax |
| 22 | BP – RP | Gaia BP-band mean magnitude—RP-band mean magnitude |
| 23 | $G - BP$ | Gaia G -band mean magnitude—BP-band mean magnitude |
| 24 | $G - RP$ | Gaia G -band mean magnitude—RP-band mean magnitude |
| 25 | $J - K$ | 2MASS $J - K$ color |
| 26 | $J - H$ | 2MASS $J - H$ color |
| 27 | $H - K_s$ | 2MASS $H - K_s$ color |
| 28 | W1 – W2 | ALLWISE W1 magnitude—ALLWISE W2 magnitude |
| 29 | W1 – W3 | ALLWISE W1 magnitude—ALLWISE W3 magnitude |
| 30 | W2 – W3 | ALLWISE W2 magnitude—ALLWISE W3 magnitude |

Data Preparation

• 2.2. Sample Set Preparation

- We built a sample set for variable source classification by **the crossmatching approach**.
- To ensure the accuracy of the sample data set labels for the machine-learning model, we selected two variable catalogs: **Chen et al. (2020) and the ASAS-SN variable star databases**. We crossmatched these two catalogs to obtain variable source labels, **selecting data with the same labels as the source of labels for our sample set**, then crossmatched them with **ZTF DR11 to obtain light-curve data**, both crossmatching with a radius of 1.5".
- Chen et al. (2020) presented a periodic variable catalog for ZTF DR2 including 781,602 variables. The 781,602 variable stars were classified into 11 classes by density-based spatial clustering with noise application.
- The ASAS-SN variable star database contains approximately 688,000 clearly labeled variable stars.

Data Preparation

- In order to improve the reliability of our research results, we have implemented the following criteria **to ensure the quality of the light curves**:
 - (1) **Select sources with at least 50 observations**. The inadequacy of having only a small number of light-curve observations introduces significant uncertainty, which hinders the possibility of conducting comprehensive follow-up research.
 - (2) **Reject the data marked as unusable**. In the study, we need to eliminate the data marked as unavailable in the light curve to ensure the reliability of the final results. Specifically, INFOBITS < 33 , 554, and 432 and catflags ≥ 32 and 768.
 - (3) **Remove data points above 3σ in the light curve** where σ is the standard deviation to eliminate occasional data point fluctuations due to inaccurate photometry.
- After crossmatching with Gaia DR3, ALLWISE, and 2MASS, we successfully **screened 44,838 variable sources with reliable labels**.

Data Preparation

- We observe an apparent imbalance between various data types.
- In order to simplify the experimental process, we **used random undersampling** to reduce the size of the EW type, which has the largest number of samples, to 8867 samples (50% of the EW samples were randomly excluded).
- After downsampling EW, our data set contains 35,972 samples, and we randomly divide the data set into a training set (70%) and a test set (30%).

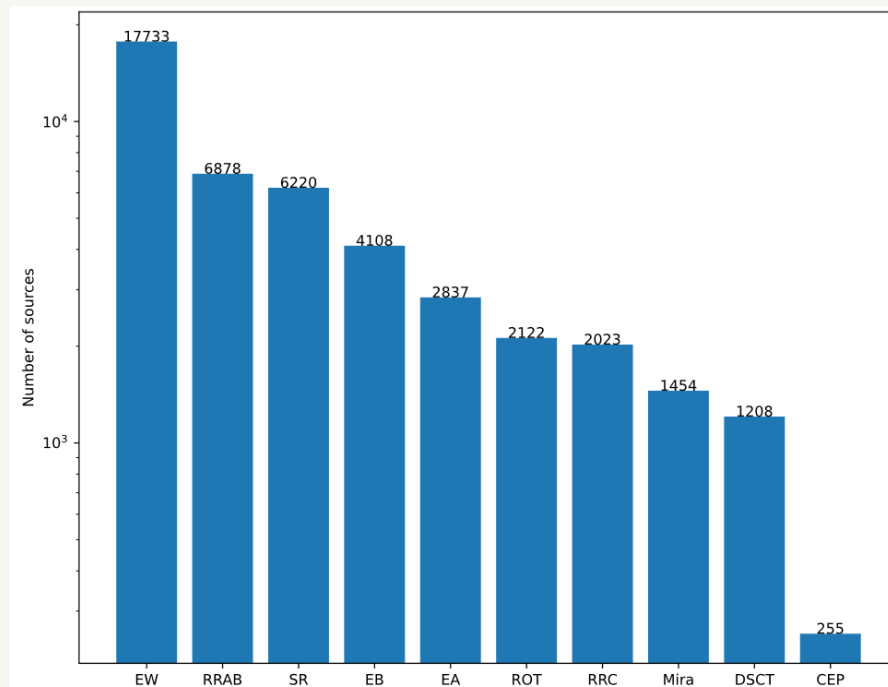


Figure 2. The histogram of the number of different variable source types in the initial sample set.

Data Preparation

- Due to the data imbalance, we also tried to introduce the **SMOTE oversampling technique** (Chawla et al. 2002). Table 2 shows the number of samples in the training and test sets in the original data set and the number of samples in the training set after SMOTE oversampling.

| Table 2 Number of Training and Test Samples in the Original and after Oversampling Using SMOTE | | | |
|-------------------------------------------------------------------------------------------------------------|--------------|--------|----------|
| Class | Training Set | | Test Set |
| | Original | SMOTE | |
| CEP | 190 | 6160 | 65 |
| DSCT | 843 | 6160 | 365 |
| EA | 2011 | 6160 | 2011 |
| EB | 2903 | 6160 | 2903 |
| EW | 6160 | 6160 | 6160 |
| Mira | 1032 | 6160 | 1032 |
| ROT | 1476 | 6160 | 1476 |
| RRAB | 4818 | 6160 | 4818 |
| RRC | 1429 | 6160 | 1429 |
| SR | 4318 | 6160 | 4318 |
| Total | 35,972 | 61,600 | 10,792 |

Classification Models

- **3.1. Modeling**

- Based on the features defined and the sample set in the previous section, we used machine-learning approaches to build classification models for turning the feature information into a probabilistic statement about the class of variable stars. We applied three machine-learning algorithms: **random forest (RF), XGBoost, and LightGBM.**

- **3.2. Model Performance**

- Machine-learning models are usually evaluated by accuracy, precision, recall, and F1 score (Forman 2003). After training and testing the RF, LightGBM, and XGBoost models, we calculated these performance metrics and performed the same evaluation on the SMOTE-augmented model.

Classification Models

Table 4

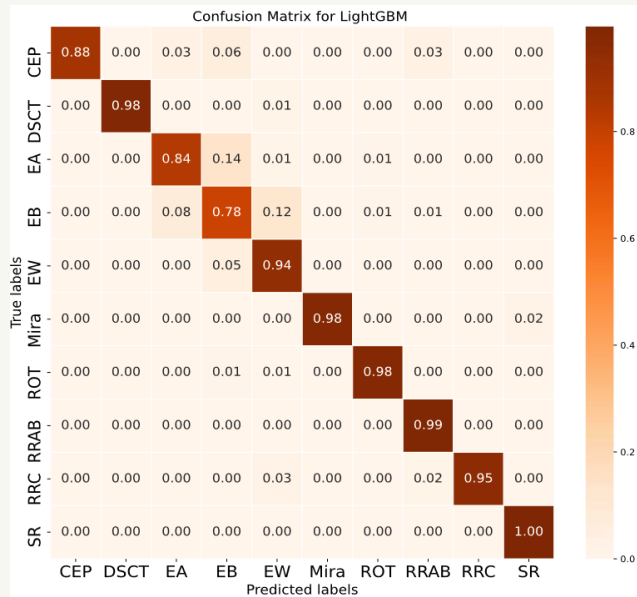
The Model Performance Metrics for 10-class Variable Stars Based on LightGBM, XGBoost, and RF with Original Data and SMOTE

| Class | LightGBM Metrics | | | XGBoost Metrics | | | RF Metrics | | |
|------------------|------------------|-----------|-----------|-----------------|-----------|-----------|------------|-----------|-----------|
| | Precision | Recall | F1 Score | Precision | Recall | F1 Score | Precision | Recall | F1 Score |
| CEP | 0.95/0.85 | 0.88/0.95 | 0.91/0.90 | 0.91/0.85 | 0.92/0.95 | 0.92/0.90 | 0.89/0.82 | 0.83/0.91 | 0.86/0.86 |
| DSCT | 0.97/0.96 | 0.98/0.99 | 0.98/0.98 | 0.97/0.96 | 0.98/0.98 | 0.98/0.97 | 0.97/0.94 | 0.97/0.98 | 0.97/0.96 |
| EA | 0.87/0.84 | 0.84/0.88 | 0.85/0.86 | 0.86/0.85 | 0.84/0.87 | 0.85/0.86 | 0.84/0.82 | 0.85/0.87 | 0.84/0.85 |
| EB | 0.78/0.77 | 0.78/0.78 | 0.78/0.78 | 0.78/0.78 | 0.78/0.79 | 0.78/0.78 | 0.76/0.76 | 0.76/0.76 | 0.76/0.76 |
| EW | 0.93/0.94 | 0.94/0.92 | 0.93/0.93 | 0.93/0.95 | 0.93/0.92 | 0.93/0.93 | 0.94/0.94 | 0.92/0.91 | 0.93/0.93 |
| Mira | 0.99/0.99 | 0.98/0.99 | 0.98/0.99 | 0.99/98 | 0.98/0.98 | 0.98/0.98 | 0.99/0.98 | 0.97/0.98 | 0.98/0.98 |
| ROT | 0.96/0.96 | 0.98/0.97 | 0.97/0.97 | 0.96/0.97 | 0.98/0.98 | 0.97/0.97 | 0.93/0.95 | 0.97/0.97 | 0.95/0.96 |
| RRAB | 0.99/0.99 | 0.99/0.99 | 0.99/0.99 | 0.99/0.99 | 0.99/0.99 | 0.99/0.99 | 0.98/0.99 | 0.99/0.99 | 0.98/0.99 |
| RRC | 0.97/0.95 | 0.95/0.97 | 0.96/0.96 | 0.96/0.94 | 0.95/0.97 | 0.96/0.95 | 0.94/0.92 | 0.93/0.95 | 0.94/0.94 |
| SR | 0.99/1.00 | 1.00/1.00 | 1.00/1.00 | 0.99/1.00 | 1.00/1.00 | 0.99/1.00 | 0.99/0.99 | 1.00/0.99 | 0.99/0.99 |
| Accuracy | ... | ... | 0.94/0.94 | ... | ... | 0.94/0.94 | ... | ... | 0.93/0.93 |
| Macro average | 0.95/0.93 | 0.94/0.94 | 0.94/0.93 | 0.93/0.93 | 0.94/0.94 | 0.94/0.93 | 0.92/0.91 | 0.92/0.93 | 0.92/0.92 |
| Weighted average | 0.94/0.94 | 0.94/0.94 | 0.94/0.94 | 0.94/0.94 | 0.94/0.94 | 0.94/0.94 | 0.93/0.93 | 0.93/0.93 | 0.93/0.93 |

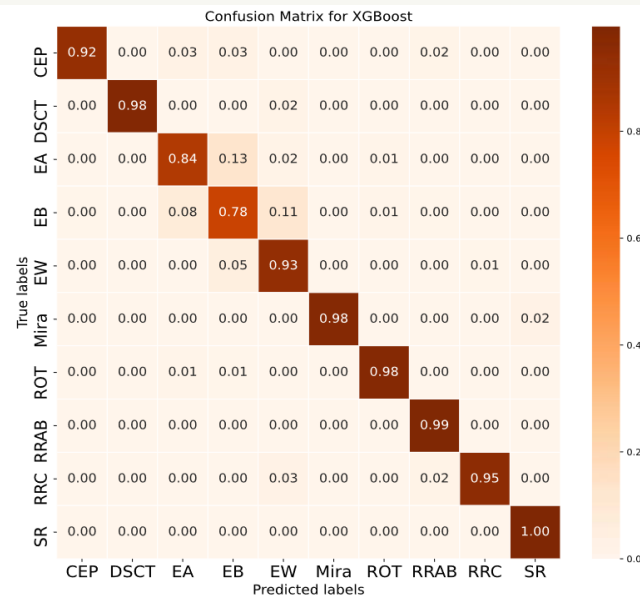
It can be seen from Table 4 that by using SMOTE for oversampling, the model improves the recall rate, but decreases the precision rate. Considering that we pursue the high precision of the final classification results, we decided to choose the model built based on the original data as the optimal model.

Classification Models

- As can be seen from the confusion matrix, our model can distinguish most of the subclasses, such as DSCT, Mira, and RRAB. However, there is some confusion between EB and EA/EW, both of which belong to the class of eclipsing binaries; their similar physical properties render them challenging to distinguish distinctly.



(a) LightGBM Confusion Matrix



(b) XGBoost Confusion Matrix

Periodic Variable Source Identification and Classification in LAMOST DR9

- After modeling the classification models, we further classified LAMOST DR9 sources.
- **4.1. Identification of Periodic Variable Stars**
- To ensure the availability of the obtained LAMOST data, we set the $S/N \geq 20$. We followed the approach presented by Xu et al. (2022) to identify the variable stars with $S/N \geq 20$ in LAMOST DR9:
- (1) Modeling variability parameters using light curves from both variable and nonvariable sources. We selected all variable sources from ASAS-SN with a classification probability greater than 95%. After crossmatching with ZTF DR11, we obtained light curves for 123,527 variable sources in the r band (observations ≥ 50). The nonvariable source labels were derived from the standard stars in SDSS (Ivezić et al. 2007), we randomly selected 123,527 light curves.

Periodic Variable Source Identification and Classification in LAMOST DR9

- (2) Obtaining the optimal model for identifying variability parameters through rigorous testing and evaluation.
- (3) Applying the optimal model to identify LAMOST variable candidates and assessing correctness through cross validation of the catalog data.
- As a result, we identified 281,514 variable sources in the r band with a confidence level exceeding 95%.
- After obtaining the variable source candidates, we searched all 281,514 variable star candidates using the Lomb–Scargle periodogram and selected periodic variable star candidates based on a false-alarm probability <0.001 , which represents the confidence of the periodicity determination. This process resulted in 198,548 periodic variable star candidates.

Periodic Variable Source Identification and Classification in LAMOST DR9

- The corresponding crossmatch results are presented in Table 5.

Table 5
The Cross Identification between Our Catalog and Published Catalogs in the r Band

| Published Catalog | Common Sources with LAMOST Unidentified Data Set | Identified by Our Model ($\alpha \geq 0.95$) | Identified Rate |
|---------------------------|--------------------------------------------------|------------------------------------------------|-----------------|
| Rimoldini et al. (2023) | 7127 | 6963 | 97.7% |
| Chen et al. (2020) | 20,845 | 19,298 | 92.6% |
| Jayasinghe et al. (2020a) | 13,116 | 12,557 | 95.7% |
| Xu et al. (2022) | 316,541 | 133,443 | 42.2% |
| Drake et al. (2014) | 7114 | 7043 | 99.0% |
| Palaversa et al. (2013) | 1753 | 1741 | 99.3% |

- We crossmatched it with Rimoldini et al.(2023) Chen et al. (2020), Jayasinghe et al. (2020a), Drake et al. (2014), and Palaversa et al. (2013), and the final identification rate was greater than 90%.
- However, our crossmatching results with the variable source catalog given by Xu et al. (2022) are relatively low.

Periodic Variable Source Identification and Classification in LAMOST DR9

- This difference mainly comes from the **different data sets**: the light curves used in Xu et al. (2022) are retrieved from **ZTF DR2**, while ours originate from **ZTF DR11**, offering a more extensive data set with increased observation points than ZTF DR2.
- Furthermore, the **labeled data set** in Xu et al. (2022) came from **Kepler**, which selects 3752 variable sources of three types: rotating variable stars, pulsating variable stars, and eclipsing binary stars. However, we chose all variable source species in **ASAS-SN**, about 120,000 variable sources, to build the variable source identification model.
- Differences in the distribution of data sets and variable source types inevitably impact the results. Overall, the results of our model in identifying variable sources are still credible.

Periodic Variable Source Identification and Classification in LAMOST DR9

• 4.2. Classification Result

- We applied LightGBM and XGBoost models to classify the 198,548 variable candidates. Given the excellent performance demonstrated by LightGBM and XGBoost, to further improve the purity of the prediction results we consider using the labels provided by the models as the final predicted classification when the predictions of the two models agree.
- Based on these two models, we generated a variable source classification catalog for LAMOST DR9 with 176,337 variable sources. Variable sources with inconsistent classification are recorded as suspected variables and not added to the catalog

Periodic Variable Source Identification and Classification in LAMOST DR9

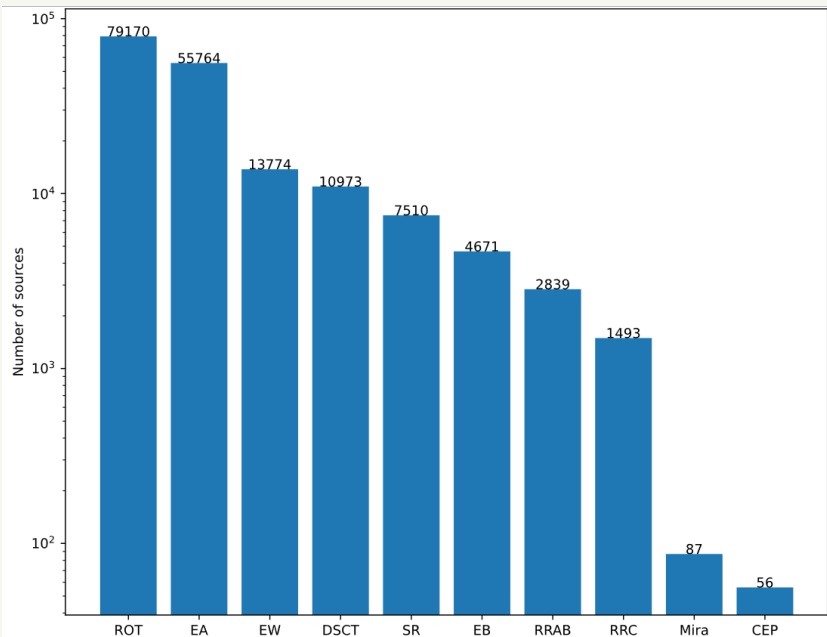


Figure 4. The histogram of predicted variable source types.

- In order to ensure the accuracy of the classification results, we adopted the method of crossmatching with the publicly available variable source catalog.
- The coincidence rate of most classes in our catalog reaches 90%.

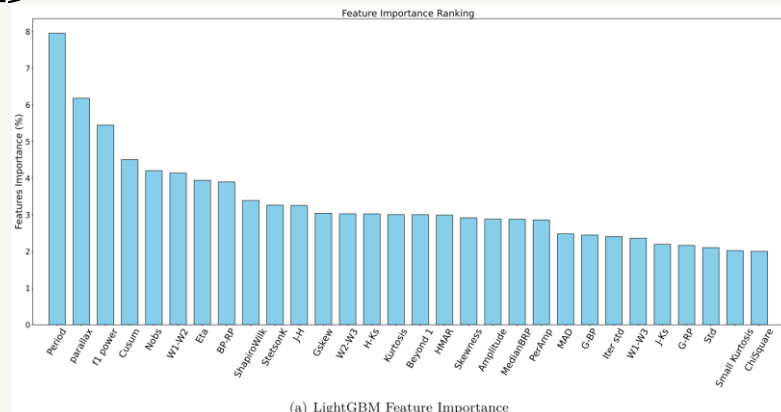
Table 7
Variable Purity Comparison between Our Catalog and Other Catalogs

| Class | Chen et al. (2020) | ASAS-SN | Gaia DR3 | Catalina | Total Samples |
|-------|-----------------------|---------|----------|----------|------------------|
| RRAB | 98.34% | 98.17% | 99.00% | 78.83% | 97.90% |
| RRC | 86.48% | 97.50% | 92.99% | 88.84% | 90.48% |
| ROT | 83.14% | 97.62% | 90.63% | ... | 90.60% |
| EW | 95.04% | 96.28% | ... | 96.77% | 95.41% |
| EA | 90.17% | 87.95% | ... | 84.16% | 87.81% |
| EB | ... | 76.14% | ... | ... | 76.14% |
| Mira | 100.0% | 92.06% | ... | ... | 96.21% |
| SR | 96.11% | 96.03% | ... | ... | 96.06% |
| DSCT | 85.88% | 91.33% | 96.87% | 58.00% | 87.85% |
| CEP | 98.04% | 100.0% | 100.0% | 100.0% | 98.82% |

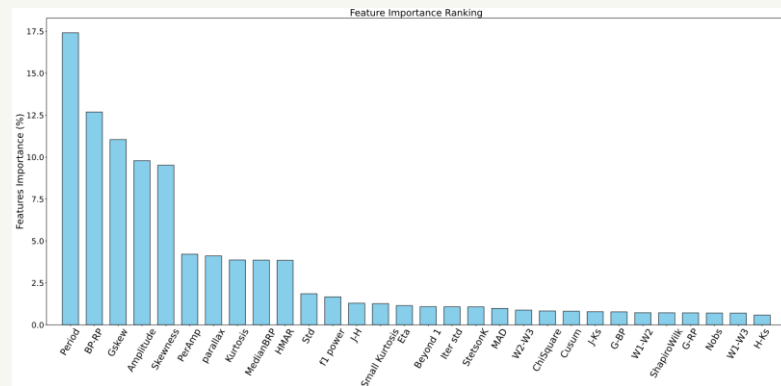
Periodic Variable Source Identification and Classification in LAMOST DR9

• 4.3. Feature Importance in the Classification Model

- To understand which features are more critical for the model's performance and decision-making process, we obtained the feature importance ranking of LightGBM and XGBoost, as shown in Figure 5.
- Period is the most significant feature in both models.
- Although the two algorithms differ in the specific ordering of feature importance, they show high consistency in evaluating results.



(a) LightGBM Feature Importance



(b) XGBoost Feature Importance

Figure 5. The feature importance of LightGBM and XGBoost.

Conclusion

- We identified candidate variable sources in the r band in LAMOST DR9 and obtained 281,514 candidate sources with probabilities greater than 95%.
- We then classified the rest of the variable sources using the XGBoost and LightGBM models, both showing high performance in our evaluation. We finally constructed a catalog of 176,337 variable sources.
- However, we must also recognize the limitations of this research. Specifically, based on the classification results of the test data, our model needs to be improved when classifying eclipsing binaries. Classifying these three subclasses of eclipsing binaries requires specialized methods, which we will consider in future work.

Thank you!