Journal of Undergraduate Research in Physics and Astronomy

RESEARCH ARTICLE | OCTOBER 01 2024

# Modern Methodologies in Machine Learning–Based Classification of Variable Stars FREE

Anjie Liu ✉; Jasmine C. Xu

Check for updates

*J. Undergrad. Res. Phys. Astron.* 34, 100002 (2024)
https://doi.org/10.1063/10.0034183

View Online

Export Citation

AIP Publishing

# Modern Methodologies in Machine Learning–Based Classification of Variable Stars

Anjie Liu[a] and Jasmine C. Xu[b]

*University of Texas at Austin, 2515 Speedway, Austin, Texas 78712, USA*

[a] Corresponding author: anjie.liu@utexas.edu
[b] jasminexu@utexas.edu

**Abstract.** Categorizing variable stars can reveal information ranging from stellar properties to cosmic distances. Traditionally, the classification process relied on human expertise and limited data. With the emergence of transformative machine learning methodologies and large-scale cosmological surveys, we are able to perform more efficient, accurate, and robust handling of expansive databases. This review traces the evolution of the classification of variable stars and highlights advancements from the latest surveys and cutting-edge technology. Reviewing papers in the field provides new insights into potential directions for improving classification methods.

## INTRODUCTION

Since their discovery in 1638, variable stars have intrigued amateur and professional astronomers alike [1]. Much is now known about variable stars—stars whose brightness fluctuates over time—including the fact that differences in factors such as period, luminosity, and mass split variable stars into several main classes and subclasses. While astronomers may be more capable of distinguishing noise from legitimate data, the feasibility of human classification of celestial objects is limited by the sharp increase in the pace and volume of astronomical data collection. Given this, it makes sense to approach astronomical questions using modern technology that can handle the vast amounts of data with reasonable efficiency and accuracy. Current research aims to determine whether machine learning models can be used to effectively classify variable stars. This is not a novel idea—many research groups have addressed variable star classification with a variety of machine learning algorithms. The most comprehensive and up-to-date surveys include the fourth phase of the Optical Gravitational Lensing Experiment (OGLE-IV), which uses microlensing to target stars in the Magellanic Clouds, as well as in the Galactic Bulge and Galactic Disk [2]. OGLE-IV provides one of the largest databases of classified variable stars worldwide, and supplies a comprehensive collection of star features, including equatorial coordinates, mean I-band and V-band magnitudes, and other photometric data. Hence, the database is well suited as a source of training data for supervised machine learning models. By building multiple classifier models, including a random forest, support vector machine, and recurrent neural network, the model that is most competent at the categorization of variable stars can be determined by computing and comparing classification metrics. However, there is still room for improvements that overcome limitations found in other studies, such as the misclassification of underrepresented variable star types.

## CLASSIFICATION OF VARIABLE STARS

Manual classification of variable stars, given vast amounts of observational data, is a tedious endeavor. Hence, it is advisable to instead produce machine learning models capable of classifying variable stars to a high degree of accuracy. Classification of a variable star is largely dependent on its light curve, which displays the brightness, or magnitude, of the star over time. This is an example of time-series data, whose feature extraction has been increasingly facilitated by the abundance of modern programming libraries such as FATS [3].

The discovery of variable stars has led to advances in determining the chemical composition of stars [4] and the distances to globular clusters and galaxies through studying period-luminosity relationships [5, 6]. Variable stars were not formally classified until 1786 by Edward Pigott, who developed a catalog of three classes: long-period variables, novae, and short-period variables, according to their light curve [7, 8, 9]. This classification was later modified and replaced by emerging and more comprehensive catalogs like the General Catalogue of Variable Stars (GCVS) as technology advanced and more accurate measures were made [10].

Although there are still ongoing studies of autoclassification of variable stars, contemporary astronomy has reached a common broad classification with some shared consensus. Generally, variable stars are classified based on their period, amplitude, spectrum, and luminosity [11]. As shown in Fig. 1, variable stars are first separated into two groups: intrinsic and extrinsic. Intrinsic stars are classified into pulsating and eruptive stars, whereas extrinsic stars are classified into eclipsing binary and rotating variable stars. Pulsating stars are divided into more specific types: Cepheids, RR Lyrae, blue large-amplitude pulsators, and long-period variables. Eruptive or cataclysmic stars are divided into supernovae, novae, recurrent novae, dwarf novae, symbiotic stars, double periodic variables, and R Coronae Borealis.

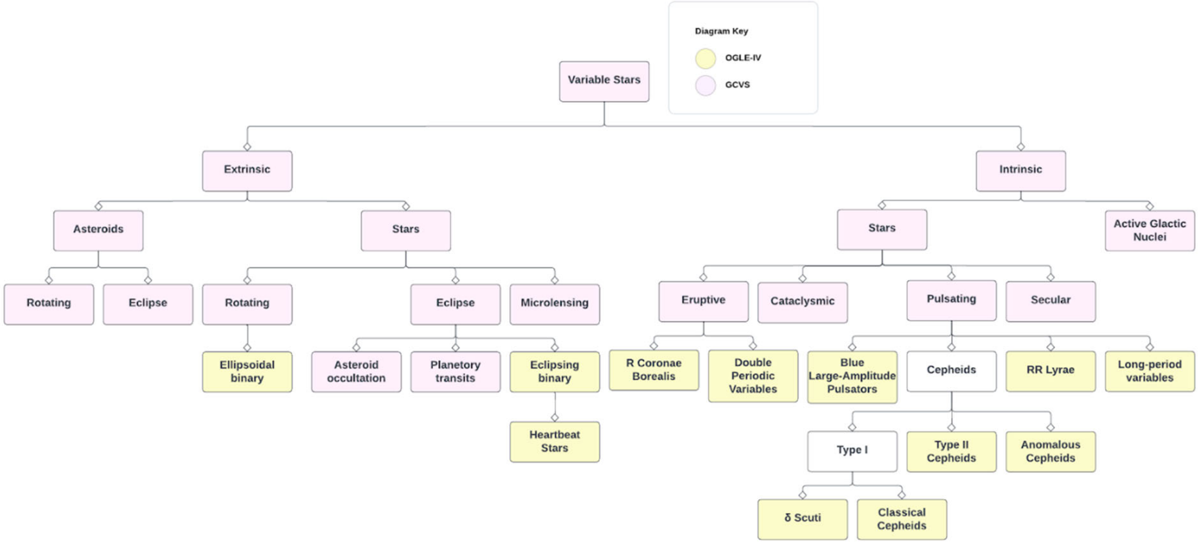

**FIGURE 1**. A simplified variable star catalog. The main branches of variable stars are classified according to the General Catalogue of Variable Stars (GCVS), which are colored pink [12]. The yellow blocks represent the variable star types observed in the Optical Gravitational Lensing Experiment (OGLE-IV) survey.

## The Optical Gravitational Lensing Experiment (OGLE)

The Optical Gravitational Lensing Experiment is one of the largest astronomical variability sky surveys in the world with an aim of discovering and classifying variable stars. The OGLE-IV phase, which began in 2011, saw the replacement of the OGLE-III phase’s eight-CCD mosaic camera with a 32-CCD camera, allowing for a greatly expanded field of view [13]. With this improvement in observing capabilities, the discovery rate of variable stars has more than doubled.

In the latest phase of the project, OGLE-IV, data are collected from four observation regions: Large Magellanic Cloud, Small Magellanic Cloud, Galactic Bulge, and Galactic Disk in the Milky Way Galaxy; they were specifically selected for having the densest star populations [13]. The observation data are collected from classical Cepheids, anomalous Cepheids, Type II Cepheids, RR Lyrae stars, long-period variables, delta Scuti stars, heartbeat stars, and eclipsing and binary stars (see Appendix). The OGLE-IV survey collected data using the large OGLE-IV 262.5-megapixel mosaic camera [13].

## MACHINE LEARNING TECHNIQUES

As can be seen from the aforementioned studies, machine learning techniques have revolutionized the advancement of various fields, including astronomy, by providing a powerful way of automating complex and time-consuming tasks such as classifying celestial objects. Traditionally, variable stars have been classified manually by astronomers who observe key patterns in star light curves [8]. As with any manual task, this is subject to biases. Automation of this task provides for a more objective and efficient approach. Moreover, the construction of such machine learning models has been vastly facilitated by the popular Python module Scikit-Learn [14].

Models aim to view each variable star as a collection of features from which patterns and relationships will be derived and used to classify unknown data. This is a subset of supervised machine learning known as classification. Classification refers to the process by which models are trained on a labeled dataset, where each input data point is paired with its corresponding correct output.

Classification encompasses several machine learning algorithms mentioned previously, including random forests, decision trees, k-nearest neighbors, and neural networks. Each of these algorithms is capable of making predictions for unlabeled data. These algorithms should be explored comprehensively, as a means of determining the most efficient model for the classification of variable stars. Classification models are not limited to the ones outlined in this section, though the mentioned handful are perhaps the most applicable and most widely used.

Random forest classifiers are a type of ensemble learning technique, in that they are collections of decision trees that seek to enhance performance accuracy [15]. Decision trees, the building blocks of random forests, simulate the decision-making process by partitioning data into smaller subsets based on features or attributes of the data. By repeating this process recursively, a classification can ultimately be made. Random forests leverage multiple decision trees and classify the input data by aggregating the collective outputs of these individual trees.

Neural networks are a broad class of deep learning techniques, of which the most popular types are recurrent neural networks (RNNs) and convolutional neural networks (CNNs). RNNs are suited for learning with sequential or time-series data, and hence would be a natural choice if raw data from variable star light curves are provided as input data. In contrast, CNNs specialize in learning from grid-like data, such as images. If the light curves were transformed into images or image-like data, as in Szklenár et al. [16], CNNs would be an ideal choice for a classification model.

One common predicament faced in the creation of classification models is hyper parameter optimization, the choosing of parameters that yield the most accurate classifications. This issue can be addressed in various ways, including by grid search, random search, or Bayesian optimization. This concern is highly relevant to future studies aiming to use the OGLE-IV database, given the sheer quantity of parameters it provides. Thankfully, resources such as the Hyperopt library [17], which is used for Bayesian optimization, exist to facilitate this process.

There are various metrics that can be used to evaluate classification models. Among the most popular are accuracy, precision, specificity (true negative rate), and sensitivity (true positive rate), which are all computed in terms of the number of true positives and true negatives. These metrics will allow us to determine how well the models are performing their task of classifying a variable star as a certain type.

## FUTURE DIRECTIONS

While the intersection of variable star classification and machine learning is undoubtedly a flourishing field, there are nonetheless ongoing challenges, as well as future directions for improvement. A common issue observed from relevant studies is class imbalance in datasets, which results in ineptitude in the model's classifications of underrepresented star types. This can potentially be addressed using data augmentation, the generation of synthetic data for these classes, or by assigning higher weights to underrepresented classes such that the model is penalized more heavily for incorrect classifications of these classes. Given the continuing generation of new data from current and upcoming sky surveys, it would be advantageous to apply machine learning techniques to these data, such that more generalizable conclusions can be drawn. Future research can address these limitations while also maintaining high accuracy.

## ACKNOWLEDGMENTS

We would like to thank Dr. Shyamal Mitra, Dr. Karl Gebhardt, Jose Ordonez, and Rik Ghosh for their support and advice throughout our research journey in the Geometry of Space stream of the Freshman Research Initiative at the University of Texas at Austin.

# REFERENCES

1. E. G. Hogg, “Mira Ceti, the ‘Wonderful Star,’” J. R. Astron. Soc. Can. **27**, 75 (1933).
2. A. Udalski, M. Szymanski, J. Kaluzny, M. Kubiak, and M. Mateo, “The optical gravitational lensing experiment,” 3. Acta Astron. **42**, 253–284 (1992).
3. I. Nun, P. Protopapas, B. Sim, M. Zhu, R. Dave, N. Castro, and K. Pichara, “FATS: Feature Analysis for Time Series,” preprint arXiv:1506.00010 [astro-ph.IM] (2015).
4. J. Govea, T. Gomez, G. W. Preston, and C. Sneden, “The chemical compositions of RR Lyrae type C variable stars,” Astrophys. J. **782**, 59 (2014).
5. P. T. Oosterhoff, “Some remarks on the variable stars in globular clusters,” Observatory **62**, 104–109 (1939).
6. G. L. H. Harris, M. Rejkuba, and W. E. Harris, “The distance to NGC 5128 (Centaurus A),” Publ. Astron. Soc. Aust. **27**, 457–462 (2013).
7. E. Pigott, “IX. Observations and remarks on those stars which the astronomers of the last century suspected to be changeable,” Philos. Trans. Soc. London **76**, 189–219 (1786).
8. E. Zsoldos, “Three early variable star catalogues,” J. Hist. Astron. **25**, 92–98 (1994).
9. J. R. Percy, *Understanding Variable Stars* (Cambridge University Press, Cambridge, 2007), oCLC: ocm81453063.
10. N. N. Samus, E. V. Kazarovets, O. V. Durlevich, N. N. Kireeva, and E. N. Pastukhova, “General catalogue of variable stars: Version GCVS 5.1,” Astron. Rep. **61**, 80–88 (2017).
11. J. A. Mattei and A. Henden, “Variable Star,” in *McGraw-Hill Dictionary of Scientific and Technical Terms* (McGraw-Hill Education, 1995).
12. L. Eyer and N. Mowlavi, “Variable stars across the observational HR diagram,” J. Phys. Conf. Ser. **118**, 012010 (2008).
13. A. Udalski, M. K. Szymański, and G. Szymański, “OGLE-IV: Fourth phase of the Optical Gravitational Lensing Experiment,” preprint arXiv:1504.05966 [astro-ph.SR] (2015).
14. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, et al., “Scikit-learn: Machine learning in Python,” preprint arXiv:1201.0490 [cs.LG].
15. L. Breiman, “Random forest,” Mach. Learn. **45**, 5–32 (2001).
16. T. Szklenár, A. Boái, D. Tarczay-Nehéz, K. Vida, G. Mező, and R. Szabó, “Variable star classification with a multiple-input neural network,” Astrophys. J. **938**, 37 (2022).
17. J. Bergstra, D. Yamins, and D. D. Cox, “Hyperopt: A Python library for optimizing the hyperparameters of machine learning algorithms,” (2013).
18. I. N. Pashchenko, K. V. Sokolovsky, and P. Gavras, “Machine learning search for variable stars,” Mon. Not. R. Astron. Soc. **475**, 2326–2343 (2018).
19. T. Fushiki, “Estimation of prediction error by using K-fold cross-validation,” Stat. Comput. **21**, 137–146 (2011).
20. G. C. Clayton, “The R Coronae Borealis stars,” Publ. Astron. Soc. Pac. **108**, 225 (1996).
21. A. Mammone, M. Turchi, and N. Cristianini, “Support vector machines,” WIREs Computational Statistics **1**, 283–289 (2009).
22. Z. Hosenie, R. J. Lyon, B. W. Stappers, and A. Mootoovaloo, “Comparing multiclass, binary, and hierarchical machine learning classification schemes for variable stars,” Mon. Not. R. Astron. Soc. **488**, 4858–4872 (2019).
23. P. F. L. Maxted and R. J. Hutcheon, “Discovery and characterisation of long-period eclipsing binary stars from *Kepler* K2 campaigns 1, 2, and 3,” Astron. Astrophys. **616**, A38 (2018).
24. B. V. Alexeev, “Nonlocal Theory of Variable Stars,” in *Nonlocal Astrophysics* (Elsevier, New York, 2017), pp. 321–377.

# APPENDIX

## Intrinsic Variable Stars

The luminosity of intrinsic variable stars changes periodically due to their physical properties [24]. Intrinsic variable stars can be divided into three main subcategories: pulsating, eruptive, and cataclysmic. Pulsation means “vibration” or “oscillation” in astronomy [9]. The surface of pulsating stars expands and contracts with a period depending on the radius, mass, and structure of the star. Cepheids are bright and massive pulsating variable stars with long periods and large amplitudes. The relationship between the period and luminosity of Cepheids is used for mapping the Milky Way Galaxy’s spiral arms and determining the distances to star clusters and nearby galaxies [9]. RR Lyrae stars and delta Scuti stars, on the other hand, have much shorter periods and smaller amplitudes as compared to Cepheids. Eruptive variables exhibit rapid and unexpected luminosity changes due to eruptions on their surfaces. R Coronae Borealis (RCB) stars are hydrogen deficient and carbon rich. Unlike other eruptive variables, RCBs experience light fluctuations by alternation of declination and recovery [20].

## Extrinsic Variable Stars

In contrast to intrinsic variable stars, extrinsic variable stars undergo an apparent fluctuation in luminosity due to external sources. Binary stars are a system of two stars orbiting around each other by gravitational pull. They provide valuable information on accurate and model-independent mass determination [23]. Eclipsing binary stars are binary stars whose orbital plane is oriented in such a way that we observe the component stars periodically passing in front of each other from Earth. Unlike eclipsing binary stars, ellipsoidal binary stars do not create eclipses as they do not pass in front of each other. In turn, both stars deform into an ellipsoidal shape due to their close proximity to each other.