Draft version September 16, 2023
Typeset using LaTeX default style in AASTeX631

Modern Methodologies in Machine Learning-Based Classification of Variable Stars

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

[1] *The University of Texas at Austin*

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 the further advancements from using the latest surveys and cutting-edge technology. By reviewing past papers in the field, it provides new insights into the potential directions for improving classification methods.

Keywords: Variable stars — Machine Learning — Astronomical surveys — Classification

1. INTRODUCTION

Since their discovery in 1638, variable stars have intrigued amateur and professional astronomers alike (Hogg 1933). 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 further subclasses. While human 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. We seek to build upon prior relevant research, and to improve upon pitfalls and limitations of their methodology. 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 (Udalski et al. 1992). 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 and comparing 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. Moreover, there are still potentials to address and overcome limitations found from other similar studies, such as the misclassification of underrepresented variable star types.

2. 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 variable stars is largely dependent on their 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, which allows for the extraction of features from time-series data (Nun et al. 2015).

The discovery of variable stars has led to the advancement of determination of the chemical composition of stars (Govea et al. 2014) and the distances to the globular clusters and galaxies by studying period-luminosity relationship (Oosterhoff 1939; Harris et al. 2013). Variable stars were not formally classified until 1786 by Edward Pigott, who

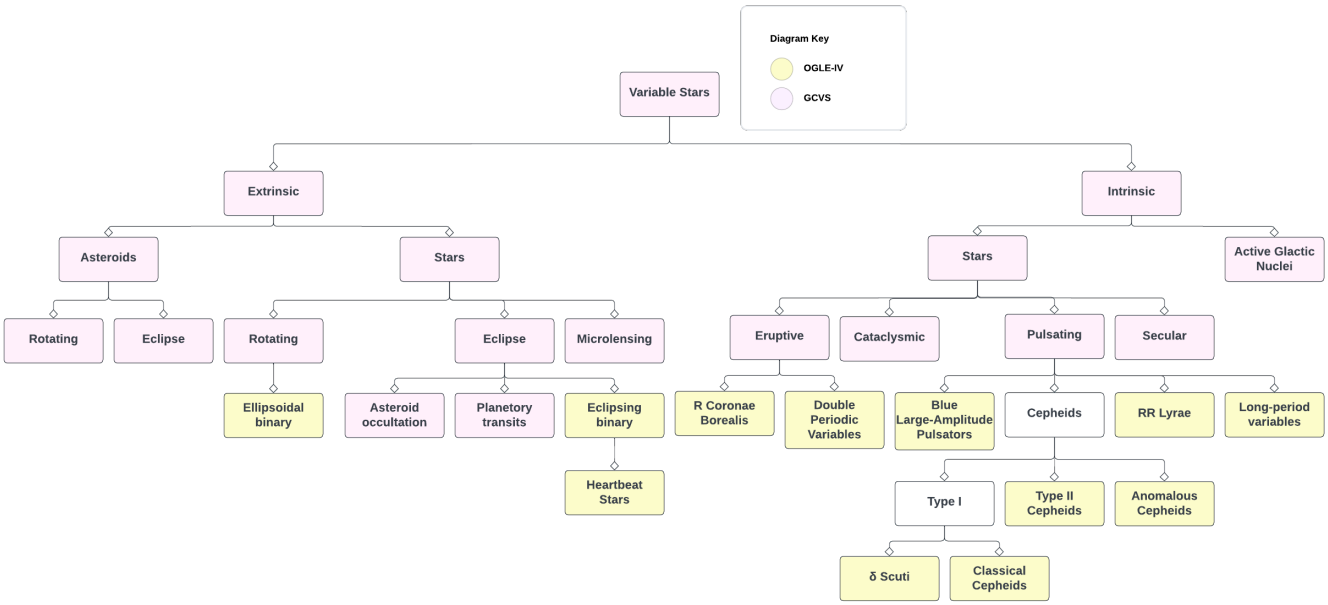

Figure 1. A simplified variable star catalog. The main branches of variable stars are classified according to the General Catalogue of Variable Stars (GCVS), which are colored pink (Eyer & Mowlavi 2008). The yellow-coded blocks represent the variable star types observed in the Optical Gravitational Lensing Experiment (OGLE-IV) survey.

developed a catalog of 3 classes: long-period variables, novae, and short-period variables according to their light curve (Pigott 1786; Zsoldos 1994; Percy 2007). This classification was later modified and replaced by new emerging and more comprehensive catalogs like the General Catalogue of Variable Stars (GCVS) as the technology advances and more accurate measures are made (Samus' et al. 2017).

Although there are still ongoing studies of auto-classification 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, spectra, and luminosity (Mattei & Henden 1995). As shown in Fig. 1, variable stars are first separated into 2 groups called intrinsic and extrinsic. Intrinsic stars are classified into 2 main classes of 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.

2.1. *Intrinsic Variable Stars*

The luminosity of intrinsic variable stars changes periodically due to their physical properties (Alexeev 2017). Intrinsic variable stars can be divided into 3 main subcategories: pulsating, eruptive, and cataclysmic. Pulsation means "vibration" or "oscillation" in astronomy (Percy 2007). 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 Milky Way Galaxy's spiral arms and determining the distances to star clusters and nearby galaxies (Percy 2007). 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, RCB experiences light fluctuations by alternation of declination and recovery (Clayton 1996).

2.2. *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 (Maxted & Hutcheon 2018).

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 on 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.

3. RELEVANT STUDIES

Prior research in the field of variable star classification forms much of the basis of the impending work. It allows us to build upon existing knowledge and avoid duplication, ensuring that the future research contributes to the field. Additionally, reviewing existing literature makes us aware of established methodologies, data sources, and potential challenges in the area of study, enabling us to make informed decisions about the research design.

Hosenie et al. (2019) created and compared the performances of random forest, decision tree, and k-nearest neighbors classifiers in the classification of variable stars. Data for this project were obtained from the Catalina Real-Time Transient Surveys (CRTS), which encompass eleven types of variable stars. They selected seven parameters that were determined to have high predictive power: I-band magnitude mean, standard deviation, skewness, kurtosis, mean-variance, period, and amplitude. These features (except for the period) were extracted from variable star light curves using the FATS package. Multi-classification of variable stars produced unsatisfactory results, with the random forest classifier achieving the highest accuracy rate of around 70%. It was found that these results could be improved by breaking the multi-class problem down to a binary classification, with the resulting random forest classifier now achieving an accuracy of over 90%. Limitations to this project include its limited scope, in that it focuses on data in the CRTS, as well as large class imbalances, caused by the prominence of some variable star types and rareness of others.

Szklenár et al. (2022) created a convolutional neural network that classified variable stars based on key characteristics of their light curves. The neural network was trained using all parameters provided by the third phase of the Optical Gravitational Lensing Experiment (OGLE-III), combined with phase-folded light curves derived from the periods and epochs from the database. Their work focused on the classification of six main variable star types: anomalous Cepheids, classical Cepheids, delta Scutis, eclipsing binaries, RR Lyrae stars, and Type II Cepheids. Each main type was then split into subclasses. This yielded high classification accuracies ranging from 77-99% for all variable star types except for anomalous Cepheids in both the OGLE-III and OGLE-IV databases. As with Hosenie et al. (2019), a limitation to this project was its inability to accurately classify underrepresented variable star types.

3.1. *The Optical Gravitational Lensing Experiment (OGLE)*

As Szklenár et al. (2022) trained their model using the OGLE-III database, it is advisable that following studies do so using the newer OGLE-IV database. There are clear advantages to using OGLE-IV over any of its older counterparts. 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 (Udalski et al. 2015). With this improvement in observing capabilities, the discovery rate of variable stars has more than doubled.

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. In the latest phase of the project, OGLE-IV, data are collected from 5 observation regions: Large Magellanic Cloud, Small Magellanic Cloud, Galactic bulge, Galactic disk, and Anomalous Cepheids in the Galaxy. 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. OGLE-IV survey collected data using the large OGLE-IV 262.5 Megapixel mosaic camera (Udalski et al. 2015).

4. MACHINE LEARNING TECHNIQUES

As can be seen from the aforementioned relevant 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 are classified manually by astronomers who observe key patterns in the star light curves (Zsoldos 1994). 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 (Pedregosa et al. 2012).

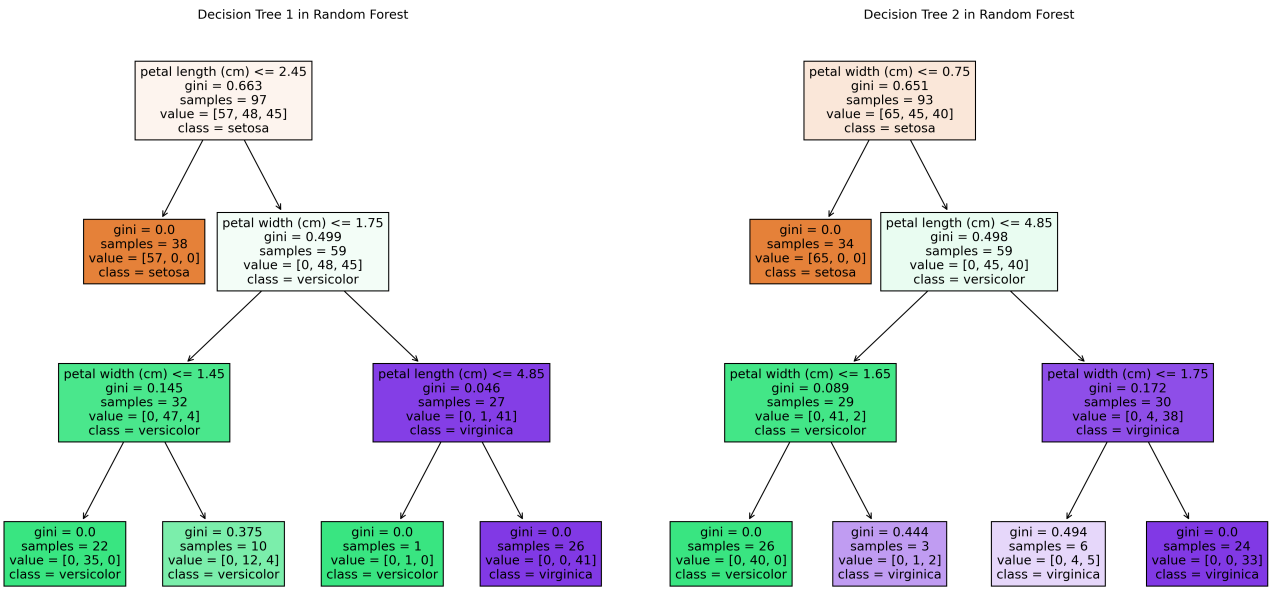

Figure 2. A simple random forest, comprised of two decision trees displayed side-by-side. This example uses the popular "iris" dataset to exemplify the branching of data that ultimately leads to a classification.

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 (Breiman 2001). 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 classifies the input data by aggregating the collective outputs of these individual trees.

The k-nearest neighbors (kNN) algorithm classifies data points using the Euclidean distance measure, under the assumption that objects of the same class exist closer to each other in space. Given a data point, evaluating the labels of its k nearest neighbors allows the model to make a reasonable prediction regarding the class of the unknown data point (Pashchenko et al. 2018).

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. (2022), CNNs would be an ideal choice for a classification model.

While not mentioned in the prior studies, support vector machines (SVMs) are yet another classification algorithm that should be considered in future research. SVMs classify objects by determining the optimal decision boundary - known as a hyperplane - in some high-dimensional space that maximizes the margin between different classes (Mammone et al. 2009). SVMs are largely dependent on support vectors, which are the data points that lie closest to the

hyperplane, and hence are the most difficult to classify. They also have a significant influence on the placement of the hyperplane, as the hyperplane aims to optimize the separation between classes.

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 (Bergstra et al. 2013), which is used for Bayesian optimization, exist to facilitate this process.

Another common issue in classification problems is overfitting, which refers to the model producing an over-optimistic result due to its over-adherence to the training data. One way in which this can be addressed is through k-fold cross-validation, which evaluates model performance on unseen data by splitting the data randomly into k groups, each of which provides a training and testing set. For each resampled dataset, called a k-fold, an evaluation score is computed, and the overall performance of the model can be determined from the collection of evaluation scores (Fushiki 2011).

There are various metrics that can be used to evaluate classification models. Of 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.

5. 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 of improvement. A common issue observed from relevant studies is class imbalance in datasets, which then results in ineptitude in model 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 accuracies in the classifications.

REFERENCES

Alexeev, B. V. 2017, in Nonlocal Astrophysics (Elsevier), 321–377, doi: 10.1016/B978-0-444-64019-2.00007-7

Bergstra, J., Yamins, D., & Cox, D. D. 2013. https://conference.scipy.org/proceedings/scipy2013/pdfs/bergstra_hyperopt.pdf

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

Clayton, G. C. 1996, Publications of the Astronomical Society of the Pacific, 108, 225, doi: 10.1086/133715

Eyer, L., & Mowlavi, N. 2008, Journal of Physics: Conference Series, 118, 012010, doi: 10.1088/1742-6596/118/1/012010

Fushiki, T. 2011, Statistics and Computing, 21, 137, doi: 10.1007/s11222-009-9153-8

Govea, J., Gomez, T., Preston, G. W., & Sneden, C. 2014, The Astrophysical Journal, 782, 59, doi: 10.1088/0004-637X/782/2/59

Harris, G. L. H., Rejkuba, M., & Harris, W. E. 2013, Publications of the Astronomical Society of Australia, 27, 457, doi: 10.1071/AS09061

Hogg, E. G. 1933, Journal of the Royal Astronomical Society of Canada, 27, 75. https://adsabs.harvard.edu/full/1933JRASC..27...75H

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

Mammone, A., Turchi, M., & Cristianini, N. 2009, WIREs Computational Statistics, 1, 283, doi: 10.1002/wics.49

Mattei, J. A., & Henden, A. 1995, in McGraw-Hill Dictionary of Scientific and Technical Terms (McGraw-Hill Education), doi: 10.1036/1097-8542.728000

Maxted, P. F. L., & Hutcheon, R. J. 2018, Astronomy & Astrophysics, 616, A38, doi: 10.1051/0004-6361/201732463

Nun, I., Protopapas, P., Sim, B., et al. 2015, doi: 10.48550/ARXIV.1506.00010

Oosterhoff, P. T. 1939, The Observatory, 62, 104. https://ui.adsabs.harvard.edu/abs/1939Obs....62..104O/abstract

Pashchenko, I. N., Sokolovsky, K. V., & Gavras, P. 2018, Monthly Notices of the Royal Astronomical Society, 475, 2326, doi: 10.1093/mnras/stx3222

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2012, doi: 10.48550/ARXIV.1201.0490

Percy, J. R. 2007, Understanding variable stars (Cambridge New York: Cambridge University Press)

Pigott, E. 1786, Philosophical Transactions of the Royal Society of London, 76, 189, doi: 10.1098/rstl.1786.0009

Samus', N. N., Kazarovets, E. V., Durlevich, O. V., Kireeva, N. N., & Pastukhova, E. N. 2017, Astronomy Reports, 61, 80, doi: 10.1134/S1063772917010085

Szklenár, T., Bódi, A., Tarczay-Nehéz, D., et al. 2022, The Astrophysical Journal, 938, 37, doi: 10.3847/1538-4357/ac8df3

Udalski, A., Szymanski, M., Kaluzny, J., Kubiak, M., & Mateo, M. 1992, Acta Astronomica, 42, 253

Udalski, A., Szymański, M. K., & Szymański, G. 2015, doi: 10.48550/ARXIV.1504.05966

Zsoldos, E. 1994, Journal for the History of Astronomy, 25, 92, doi: 10.1177/002182869402500202