MNRAS 000, 1–6 (2022) Preprint 15 August 2022 Compiled using MNRAS LATEX style file v3.0

Evolution in Pulsating Variable Stars: Long Term, and Inter-Cycle

Uzair A. Khan,[1]
[1] University of Hull, Cottingham Rd, Hull HU6 7RX, UK

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We used machine learning techniques and built models to predict when variable stars might change their periods using noisy and sparse time series data and also by inference to learn about the underlying physics of these stars and to predict not only the long term evolution of such stars, but to try to see if we can generalize predictions about cycle-to-cycle variations. For this purpose we chose Mira variables which is a well known class of pulsating variable stars, we pre-processed data point of four stars from Mira variables namely Mira, R Andromedae ,U Orionis and Chi Cygni available at American Association of Variable Star Observers (AAVSO) to predict luminosity magnitude uncertainty and classify pulsation state using a selection of various classification and regression algorithms also utilizing feature engineering methods, since one of the primary goal is to generalize predictions.

We use a data-set that shows the mentioned stars data points and fix our scope for machine learning for a common time duration, hence creating a generalized data-set with collective averaged data points. Linear regression models with single and multi input produced no successful predictions but Decision Tree and KNN regressors proved successful in predicting luminosity magnitude errors which is the amount of variation observed over a repeated measurement over a time frame. Feature engineering was successful for both regression and classification of the state of pulsating stars.

In particular, we were able to achieve classification accuracy of 0.8 after hyperparameter tuning using Bayesian Neural networks and by KNN classifier classification accuracy of 0.94 for the same classification of pulsation state of Mira variable stars with a R^2 score of 0.98 from the regression model. This work has the potential to provide a foundation for developing tools to further aid in the analysis of various pulsating star variables like Cepheids variables, RR lyrae variables, Delta Scuti variables and other astrophysical data for classification and regression and performs impressively with time series datasets.

1 INTRODUCTION

Studying the Milky Way Galaxy generates large amount of data points and there is a huge potential in future to add even larger amount of unprocessed data into the astronomical community to explore, an important reason to process these datapoint is to look into the behavior and structure of similar galaxies, to find patterns and predict the evolution of stars s similar study was conducted where a feature, period of Mira variables were determined using machine learning where the string length method combined with a least-square fit curve selection proves to be an promising method to extract periods of Miras from the VVV sky survey [9]. It has been a challenge to spatially resolve and predict the motion of pulsating stars as they expand and contract and have cycle to cycle variations causing them to pulsate and change their size and flux rhythmically.

Variable stars are systematically observed over decades to determine their behavior and detect common patterns present within the pulsating variables, due to limitations in availability of time and storing of large volumes of recorded data of thousands of variable stars the process of investigating them is quite a challenge and a even bigger challenge is to develop machine learning methods that can highlight the patterns and predict certain features that variable stars inherit. Therefore, amateur astronomers observing these stars make a highly useful contribution to the astronomical community and science by submitting observations to the American Association of Variable Star Observer's (AAVSO's) International database (ID). These submissions help analyze variable stars and to make computerized algorithms for their better predictions and classifications. Utilizing the approaches of machine learning we started the analysis by applying feature engineering and taking raw unprocessed data from the data-sets and summarizing them with introducing features deemed to be important, then the newly formed features are fed into machine learning algorithms for classifications and regression purposes.

The work in this analysis is divided in two separate parts, an approach for classification after feature engineering and hyperparameter tuning using Bayesian Theorem then feeding the configuration and feature engineered averaged data-set to the Neural Networks and the second part involves generating predictions using regression models. To the best of our knowledge this is the first work to do the predictions for differences in variations of luminosity Magnitude of variable stars (Uncertainty) and classification on state of pulsating variable stars (quasi-equilibrium or stable), however a notable work was conducted where machine learning techniques used to predict periods of Mira variables pulsation using wavelet analysis to analyze evolution in pulsating stars [4].

2 METHOD

As stated in the introduction we will discuss in detail the approach taken for this analysis . In this section we discuss how the method was implemented . Results and findings will be presented in the subsequent sections.

2.1 Selecting Data-sets

In an attempt to make this study credible and acceptable , the data-sets selected from pulsating Mira variable stars are chosen at random and for this analysis the chosen data-sets are Mira , R Andromedae , U Orionis and Chi Cygni . The raw data-sets of the selected stars have flux differences with different magnitudes , phases , periods and cycle to cycle fluctuations . This diversified approach is an effective attempt for machine learning , upon looking inside the datasets most columns are completely filled with data but the Uncertainty which is the error in magnitude luminosity of pulsating stars recorded seems to be sparsely completed.

2.2 Preprocessing Data-sets

Data-sets of Mira variable stars are acquired from AAVSO, an organization who allows anyone to participate in scientific discovery by making the records of variable stars available to the public . The AAVSO repository allows photometric data-sets to be extracted to create light curves for further analysis .

Before training the remaining data , a number of preprocessing steps are taken . These are given in order below

1. Dataframe columns are restricted to “JD” , “Magnitude” and “Uncertainty” .
2. Special characters present within the columns i.e “<” , “>” attached with float values that restrict the data type to be object rather than float type are removed .
3. Identify columns in the data-sets with consecutive missing values denoted with NaNs = “Not a Number”
4. The Magnitude (Flux) data points in the data-sets are observed over a period of time recorded in Julian Date (JD) while also giving the Uncertainty (Error) of the observations recorded. The data-sets showed that column Uncertainty has a large volume in data points missing , so in an attempt to make the models performance better. Therefore It was decided to impute data by splitting the data into groups based on the columns with completely filled real values, in this case “Magnitude”, after grouping the missing data points data was imputed by introducing linear interpolation hence replacing the missing data with grouped interpolated data points between the two real values, which can be seen in the referenced code provided in the code log book.

2.3 Discrete data points

Before feeding the data points of the Mira variables to the algorithm it was necessary to look into the data points and find areas where discrete data exists, which was found after plotting the data sets in where x axis defines Julian Date and Y axis defines the Magnitude of luminosity *Figure 1*.

The light curve displayed presents data points of the four Mira stars showing higher concentration of data observations recorded after 2420000 Julian date and further on wards in timeline, whereas the data points plotted before the aforementioned range is quite sparse and might affect the performance of our machine learning models, in order to maintain consistent data, the discrete data is removed from the data set. This is a consistent common practice used within the scientific community. This reduces the noise and sparse data and optimizes machine learning *Figure 2*.

Therefore constructing the light curves of the filtered data presents a more rich and continuous sinusoidal wave as shown visually proven in the light curve hence thinning this will prove to be an effective attempt at machine learning models.

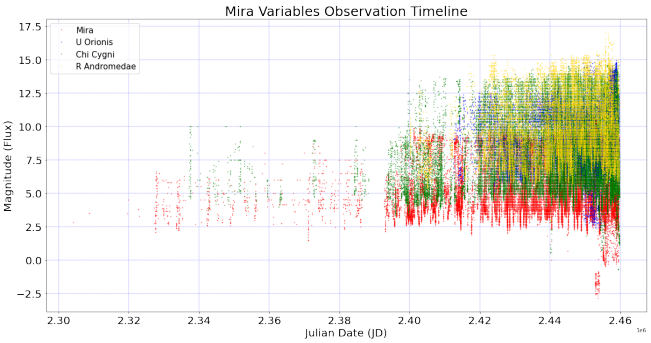

Figure 1. The raw time series of Mira , U Orionis , Chi Cygni , R Andromedae demonstrating light curves of the Mira variable stars.

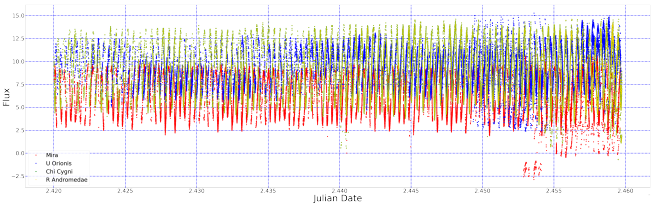

Figure 2. The filtered time series data of the Mira variable stars tailored to demonstrate rich and continuous light curves

2.4 Feature Engineering

2.4.1 Phases and Cycles

If a star is periodic, then the variation depends upon its position in its cycle which is called a phase. However, phase is measured in cycles which is usually a fraction of the cycle that lies in the range of 0 to 1, as mentioned above the variation is dependent on its phase it is important for the analysis. Phases can be be calculated by finding the difference between the time of the start of cycle to and at any other time t.

$$t - t_0$$

Further to get phase in units of cycles, we divide this with the period of individual stars. Here the decimal part of the cycle is its phase.

$$\phi = \frac{(t - t_0)}{P}$$

$$\phi = decimal part of [\frac{(t - t_0)}{P}]$$

Using the above mentioned approach we calculated the total cycles of Mira, R Andromedae, U Orionis and Chi Cygni up to the present date using their periods for a complete cycle 332, 372.4, 408 and 409 days respectively and plotted their cycles on timeline recorded on x axis showing their spread of cycles *Figure 3*.

Then taking the decimal part of the cycle and repeating the process for all the stars to get the standard phase diagrams and plotting them to exhibit their phase differences *Figure 4*.

2.5 Quasi Equilibrium State / Pulsating state classification

Since every pulsating star exhibits a different phase behavior it is important to have a generalized phase diagram of the selected Mira

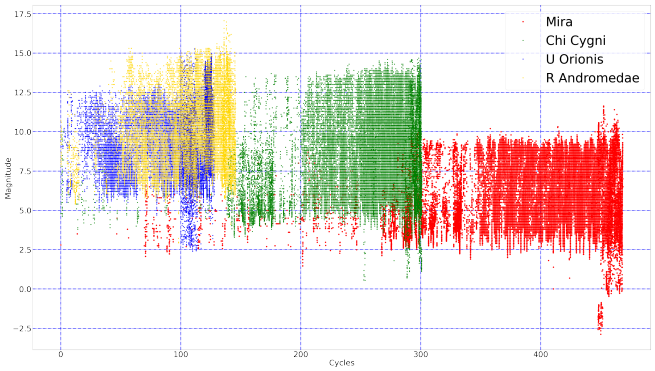

Figure 3. The figure demonstrates the cycle spread of the focused variable stars also highlighting the cycle to cycle fluctuations of the pulsating stars, The period of time decide the spread of the cycles, where Mira stars has the lowest cycle period of 332 hence the longest cyclic spread and R Andromedae exhibiting the highest cycle period of 409 hence the smallest cyclic spread.

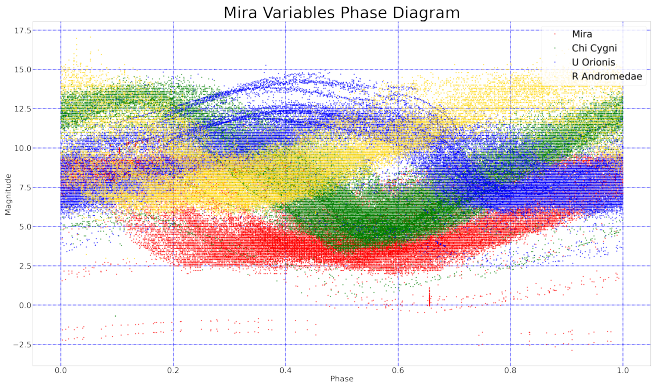

Figure 4. Plotting all the individual standard phase diagrams in a single plot provides us with the magnitude flux peaks of our model $2.5 \leq R \geq 15$ and which will in further steps assist us to make generalized light curves to feed into our machine learning models.

variable so a more widely applicable ML algorithm can be obtained. For this purpose we classify 0.9 and 0.1 percentile of the Magnitude spread as "1" denoting brightness or dimmest state of the star and the rest points as "0" denoting the regular pulsation state of the star also known as the quasi equilibrium state. The 0.1 and 0.9 percentile of the luminosity Magnitude of stars is selected after iterations of random selections to achieve best possible accuracy when feeding it to the ML algorithms.

2.6 Generalized Data Points

We approach the stage where we need to craft the data points that would be fed to the machine learning algorithms and for that reason we need generalized data points that are in common between all the light curves of the data sets of Mira variable chosen for this study by merging the data sets using common Julian dates . We now implement a common approach of averaging the data points of each column in the data set to construct a more applicable data set that could be used for ML learning *Figure 5*.

Once again we apply the feature engineering to find the cycles and phase of this new general light wave by taking the Period of average of all Mira variable stars and then within the defined Julian date time.

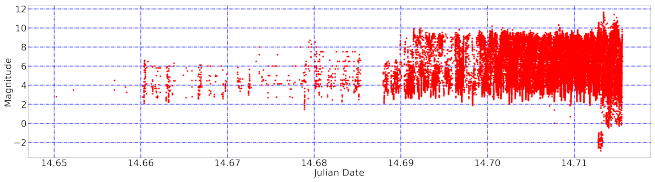

Figure 5. The light curve generated from the Mira variables inner merging on the ground of the common timelines.

Table 1. Binary classification of the pulsation state of Mira variables

Labels	Binary Value
Pulsating State	1
Not Pulsating State	0

2.7 Pulsation State Classification

Two different models were tested for the pulsation classification for the Mira variable stars to obtain the best classification accuracy for the pulsation state.

2.7.1 Bayesian Neural Networks

Hyper parameter tuning is an essential part of the overall analysis as it is responsible for finding the optimal configuration to enhance our machine learning models performance and decides the efficiency of its problem solving. For this purpose we used Bayesian Optimization method method by feeding it the generalized data points to find the optimal configuration .

This method works by developing a posterior distribution of functions that best describes the function we want to optimize. As the magnitude of observations grows, the posterior distribution improves, and the algorithm becomes more certain of which areas in parameter space are worth exploring and which are to be dropped out of scope. As you iterate over and over, the algorithm balances its needs of exploration and exploitation taking into account what it knows about the target function.

$$P(f|D) = P(D|f) * P(f)$$

Posterior Probability P(f|D) - The conditional probability that we are calculating is referred to generally as the posterior probability.

Likelihood P(D|f) - The reverse conditional probability is sometimes referred to as the likelihood.Prior Probability

P(f) - and the marginal probability is referred to as the prior probability.

Using this approach the parameters extracted from Bayesian optimization for optimal results are loaded in the Neural Net model for classification of pulsation state whether the star exhibits quasi equilibrium state or not *Table 1*.

The generalized dataset pulsation is label encoded into binary values for classification using the Bayesian neural network , obtaining a cross validated accuracy of $\sim$ 80% for classification of pulsation state.

2.7.2 K-neighbors Classification

For this classification of pulsation state of Mira stars we use a supervised learning model well known for its use of proximity for

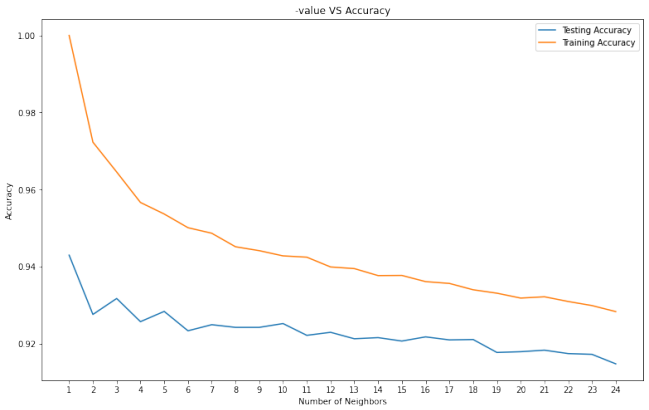

Figure 6. Performance of KNN classifier based on the minimum neighbors for maximum accuracy for classification of Pulsation state.

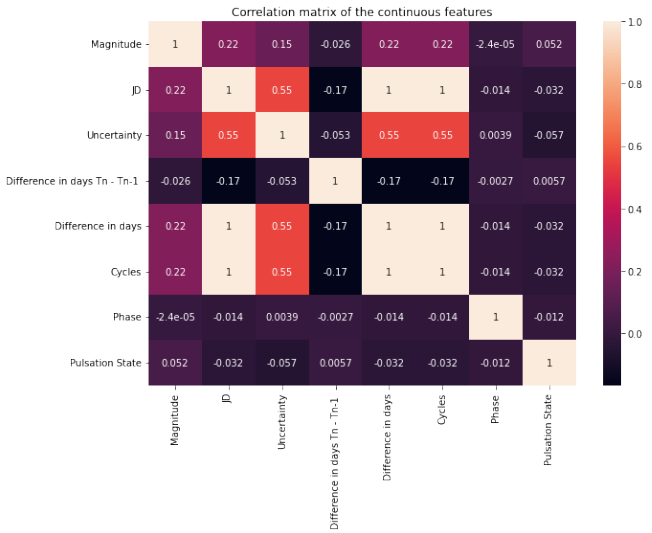

Figure 7. Correlation Matrix showing the most correlated variables for Uncertainty for input in Regression models.

classification, using the same generalized data-set without label encoding we feed the data points to the KNN classifier with 1 nearest neighbor giving highest classification accuracy *Figure 6*.

The final accuracy of the KNN classifier resulted in an accuracy score of 0.94 (~ 94%) in classification of pulsating state of Mira variable stars.

2.8 Predicting Uncertainty in Mira Variable

The variation of luminosity is continuous in pulsating variable stars, as it contract and expands producing changes in temperature and radius, Flux measurements are recorded in the data-set contains a level of error from the true value making the measurements less accurate, if that uncertainty can be predicted we can reach to the true value by identifying and predicting the value of error. To predict the Uncertainty (error) of the observation present in the data-set, we approached this by experimenting with four different regression models to test and predict Uncertainty. Before we move to the regression models it was important to identify the data point classes that correlated with the continuous feature of Uncertainty, for this purpose we create a correlation matrix to identify correlation coefficients between the variables that can be seen in the *Figure 7*.

This correlation matrix helps us identify columns that are more correlated to the Uncertainty error which we will use as input to the regression models to predict the uncertainty.

A linear regression model uses relationships between the data points to identify linear relationships drawing a straight line and hence predict future values. We configure the dependent variable as the Uncertainty against the independent variable Julian Date (JD), using this single feature we will predict the values find the coefficient of determination R^2. Similarly we also configured a multi variable regression model to predict Uncertainty and find the goodness of fit R^2, both of the score for the models were not good compared to the R^2 obtained from the Decision Tree Regression model and KNN regression models. A decision tree is used to fit curves with addition noisy measurements by breaking down the data-set into smaller subsets, the model can be used for regression analysis and hence predict the uncertainty for this part, we used this approach to check the prediction and R square score for the predictive model after feeding it the generalized data-set. The prediction to Actual value of decision Tree regression can be seen at *Figure 8*.

In the next step in search for better performance of the ML algo-

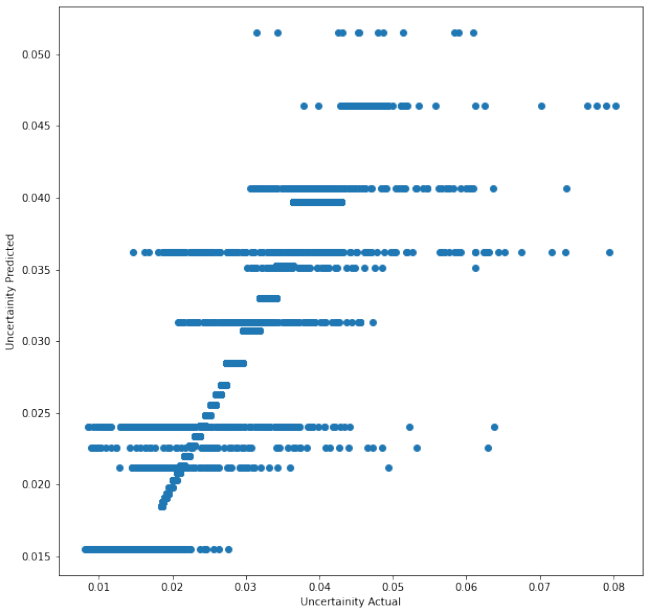

Figure 8. Actual vs Predicted relationship performance of Decision Tree Regression.

rithm for predicting values we tried K-neighbor Regressor that used a local average approach to predict the values.

3 RESULTS

While the results from Bayesian neural network for pulsation state classification and both linear regression for prediction of Uncertainty values were initially disappointing , this led us to investigate and feature engineering for these models , hence further investigation led us to find very promising results with both classification and regression accuracies crossing the threshold set by the previous failed approaches.

The classification accuracy in line with other results achieved

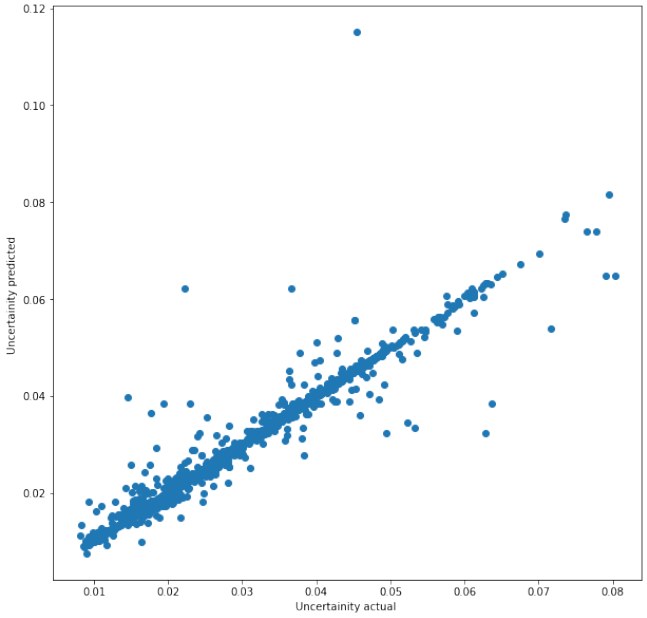

Figure 9. Actual vs Predicted relationship performance of KNN Regressor.

Table 2. Performance of binary classification models of Mira variable stars

Classification Model	**Performance Accuracy**
Bayesian Neural Network	78%
KNeighbors Classifier	94%

$\sim 90-95\%$, although Noisiness and sparseness of the Mira data in the Uncertainty (error) of the Magnitude Luminosity appeared to play a big role to classify light curve of the generalized data-set with the machine learning methods used. The classification accuracy can be compared with the machine learning approach to classification of variable stars using Cepheid group using SMOTE, selecting a minority class instance A at random and finding its k-nearest minority class neighbors B at random and connecting A and B to form a line segment in the feature space from (Naydenkin et al., 2020) achieving a classification accuracy of 81.4% whereas classification accuracy using Mira variable stars used in our study achieved 94%, the performance accuracies of our models can be seen in the *Table 2*.

Initially we assumed that the time series data-set of Mira variables fed to neural networks were not suited to be tested with machine learning algorithms but as we tested with different local averaging classifiers the predictive performance accuracy of the model significantly improved from 78% towards 94%.

Feature engineering provided important features that were used for the prediction regression purposes of the uncertainty values, although R^2 scores of the first two models weren't satisfactory after testing by cross validation of test data, but further investigation and experimentation with other machine learning models like decision tree that creates a subset of data-set and test and KNeighbor regression model gave significantly promising results compared to the initial regression models tested. The prediction performance of the regression models can be seen in *Table 3*.

Linear regression and Multi variable linear regression models might not be successful in predicting time series data with sparse and variable data points as per the nature of Mira variable pulsating stars but KNN regressors and decision trees have shown promising performance in prediction the output sample of the predicted values from the successful regressor can be seen in the *Table 4*.

Table 3. Regression models performance in predicting values of error in magnitude flux recorded

Regression Models	R^2
Linear Regression	0.2994
Multi variable Linear Regression	0.3010
Decision Tree Regressor	0.8426
KNN Regressor	0.9794

Table 4. Comparison of the cross validated regression values with the actual values.

True Values	KNN prediction	Decision Tree Prediction
0.031911	0.031913	0.033014
0.028071	0.028073	0.028503
0.022623	0.022727	0.031323
0.018500	0.018500	0.018500
0.018500	0.018500	0.018500

4 SUMMARY

With the recent advances in data science and machine learning, new methods and learning are being developed to counter the eminent boom of astronomical data, most importantly variable pulsating stars thus they are reducing the analysis time and increasing performance of machine learning algorithms in predictions and classifications, while also providing insights into the data.

Upon investigation and deep diving into the variable pulsating stars datasets available at AAVSO we found techniques to apply machine learning algorithms that could help astronomers in predicting values and classifying of pulsation states of variable stars, we chose Mira pulsating variables for this purpose. The datasets used for the experiment were utilized to construct light waves and some necessary filters were applied to achieve the desired results. The data was sparse and had missing data columns which were dealt with by the approach of grouped by linear interpolation between the two real values and proved successful in bringing the modifications required as the KNN classifier classification accuracy of 94% reached. The feature engineering part provided excellent results that further enhanced the performance of the regression models hence providing a prediction accuracy results of 98%. The same regression models can be used to predict other properties of pulsating variable stars using the same local average technique as there are cycle to cycle differences and fluctuations in a pulsating star. Which we have seen in our analysis as well.

Upon inspection of the literature in the astronomical community this may be the first study on the pulsation state of variable stars using machine learning methods and utilizing data-sets from the AAVSO database. We hope this work will provide the foundation and starting

step for the future development of tools and techniques to study other pulsating variable types of stars.

REFERENCES

[1] SA Zhevakin. "Physical basis of the pulsation theory of variable stars". In: *Annual Review of Astronomy and Astrophysics* 1 (1963), p. 367.

[2] JR Donnison and IP Williams. "Luminosity and temperature relationships for extrasolar planets". In: *Monthly Notices of the Royal Astronomical Society* 325.4 (2001), pp. 1497–1499.

[3] Roberto H. Méndez. *Light Curves of Variable Stars*. 2005.

[4] MR Templeton, JA Mattei, and LA Willson. "Secular evolution in Mira variable pulsations". In: *The Astronomical Journal* 130.2 (2005), p. 776.

[5] Donna L Young. "Variable Star Astronomy Education & Public Outreach Initiative". In: *Society for Astronomical Sciences Annual Symposium*. Vol. 27. 2008, p. 87.

[6] Hilding R Neilson, John R Percy, and Horace A Smith. "Period Changes and Evolution in Pulsating Variable Stars". In: *arXiv preprint arXiv:1611.03030* (2016).

[7] Bradley W Carroll and Dale A Ostlie. *An introduction to modern astrophysics*. Cambridge University Press, 2017.

[8] Trisha A Hinners, Kevin Tat, and Rachel Thorp. "Machine learning techniques for stellar light curve classification". In: *The Astronomical Journal* 156.1 (2018), p. 7.

[9] Kylar Greene. "Determining periods of Mira Variables using the VVV sky survey". 2019.

[10] Kirill Naydenkin, Konstantin Malanchev, and Maria Pruzhinskaya. "Variable Stars Classification with the Help of Machine Learning". In: *CEUR Workshop Proceedings*. 2020, pp. 289–296.

[11] Vıctor Muñoz and N Elizabeth Garcés. "Analysis of pulsating variable stars using the visibility graph algorithm". In: *Plos one* 16.11 (2021), e0259735.