This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Long Period Variables in the VVV Data

A. Bradley

Doctor of Philosophy
The University of Edinburgh
July 2022
Abstract

The variable stars known as Ogle Small Amplitude Red Giants (OSARGs) are potentially one of the most numerous types of variable star but relatively little is known about them. OSARGs and other Long Period Variables (LPVs) are evolved low mass pulsating variable stars. The properties and variability of these objects in the NIR compared to the optical is examined. They are much more common than other pulsating variable stars such as Cepheids, whilst also obeying a Leavitt law styled period-luminosity relation. To aid in the calculation of such a relation at the near-IR, I have compiled a catalogue of these objects. The variables in the catalogue were detected using two separate machine learning networks, trained on the OGLE catalogues of red giant stars.

The result is a catalogue of $\sim 34,000$ eclipsing binaries, $\sim 48,000$ LPVs and $\sim 25,000$ other pulsating variables. The periods calculated for these LPVs suffered due to a lack of epochs of observation, leading to a lack of a well-defined period-luminosity relation.

Unexpected behavior of the OSARGs in the Ks band was noticed, with them having $\sim 5x$ higher amplitudes in Ks than their OGLE counterparts in I band. This goes against the pattern fit by other pulsating variables including Cepheids and Mira, where the amplitude of variation decreases at higher wavelengths. A number of observational or systematic explanations were examined but none fit the observed behavior. My current theory is some facet of the OSARGs outer layers dampens the variation amplitude in optical wavelengths, but this cannot be confirmed without further data.
Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(A. Bradley, July 2022)
Acknowledgements

To Mum, Dad, Nerida and Blair,

To my supervisor Nick Cross,

To Andy Lawrence, Carlos Ferreira Lopes, Olivia Jones and everyone else at the RoE and beyond who helped me throughout my thesis.

To my friends, Bob, Cin, Alex, Andrew, other Alex Craig, Mark, Monica, other other Alex, Eddie, Peter and, well, everyone.

I don’t think I’d have completed this without everyone so I think this entitles you to about $\frac{1}{20}$th of a thesis? Good job all round.

I will be grateful for every one of you as long as I live.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>i</td>
</tr>
<tr>
<td>Declaration</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>v</td>
</tr>
<tr>
<td>Contents</td>
<td>vii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xxiii</td>
</tr>
</tbody>
</table>

## 1 Background

1.1 Sources Used and Other Surveys .......................................................... 4

1.2 Variable Stars ........................................................................ 11

   1.2.1 Intrinsic: Pulsating .......................................................... 12

   1.2.2 Extrinsic: Eclipsing .......................................................... 22

   1.2.3 Extrinsic: Rotating .......................................................... 25

   1.2.4 Intrinsic: Erupting ........................................................... 27

   1.2.5 Intrinsic: Cataclysmic ....................................................... 30

   1.2.6 Intrinsic: X-Ray Bright ..................................................... 32
2 Variable stars and their properties 35

2.1 Targets ................................................................................. 36
  2.1.1 Target Details ........................................................................ 38
  2.1.2 Additional complications ................................................... 44

2.2 VVV Observations ................................................................. 53
  2.2.1 Known Issues ...................................................................... 55
  2.2.2 Crossmatching .................................................................. 56

2.3 Statistical Properties .............................................................. 59
  2.3.1 Temperatures ....................................................................... 61
  2.3.2 Periods ............................................................................. 69
  2.3.3 Improving Period Finding ................................................... 76

2.4 OSARG Specific Issues ........................................................... 79
  2.4.1 Effects at other wavelengths .............................................. 85
  2.4.2 OSARG Simulations ......................................................... 87

3 Machine learning for variable star classification 97

3.1 Introduction ............................................................................ 97

3.2 Selection of Candidates .......................................................... 99
  3.2.1 Training Data ...................................................................... 99
  3.2.2 NIR and Optical Properties ............................................... 102
  3.2.3 Selection of LPVs ............................................................. 103
  3.2.4 Variability Indices ............................................................ 123

3.3 Probabilistic Random Forest (PRF) ........................................... 134
  3.3.1 Implementation ............................................................... 139
List of Figures

1.1 Filter efficiency of the VVV filters, with the quantum efficiency of a sample detector plotter in pink for comparison. Filter properties image from the Cambridge Astronomy Survey Unit website (Unit). ................................................................. 5

1.2 Field of View of several infrared camera, comparing and demonstrating VISTA’s particularly wide field. The moon is also included for reference (VISTA/ESO, VISTA/ESO). ......................................................... 6

1.3 A comparison of the decrease in magnitude one would observe through equivalent amounts of extinction between optical and near-IR. The dotted and dashed lines denote different extinction laws; a simple power law for the filled lines and Cardelli et al. (1989)’s law of extinction for the dashed lines. The grey and black colours are for Gaia G band and V band respectively. Figure 2 from Matsunaga (2017). ...................................................................................... 7

1.4 The maximum extinction at which a 10 day period cepheid can be detected with a given limited magnitude. The red area and dotted line show a real example case, with extinction (and error bar) increasing with respect to depth, drawn from a 3D extinction map, whereas the dot-dash line indicates the extinction increasing at a constant rate. The dotted and dashed lines denote different extinction laws; a power law similar to Nishiyama et al. (2006)’s law of extinction for the filled lines and Cardelli et al. (1989)’s law of extinction for the dashed lines. Figure 3 from Matsunaga (2017). ...... 8
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>A figure of the evolutionary position of the pulsating variable stars on the HR diagram. Distribution of different types of pulsating variable stars across the HR diagram. The solid black line is the zero-age main sequence (ZAMS). The dotted lines labelled with numbers represent evolutionary tracks for a star of that mass. The dot and dash line is the zero-age horizontal branch. The dotted lines to the left are the white dwarf cooling curve for He and CO white dwarfs. The different shading patterns represent the different pulsation mechanisms - diagonally to the left are p-mode, diagonally to the right are g-mode pulsations. Vertical lines are non-adiabatic pulsations, horizontal lines are solar-like pulsations. Figure originally from Jeffery (2008), recreated by Catelan &amp; Smith (2015a).</td>
</tr>
<tr>
<td>1.6</td>
<td>Variability tree, separating the different types of variable objects according to the cause of variation. The starting divide is between internal phenomenon occurring in the star and those from outside effects. Figure 1 from Gaia Collaboration et al. (2018b).</td>
</tr>
<tr>
<td>1.7</td>
<td>Cepheid CD Cyg, showing its variation in temperature (left) and log gravity (right) over one cycle. The red points denote measured values whereas the blue line denotes calculated values. Authors note that the large log g variations are probably due to bad decomposition of the line profiles used to measure the gravity. Figure 1 from Rastorguev et al. (2019).</td>
</tr>
<tr>
<td>1.8</td>
<td>Period-Wesenheit (a reddening-free quantity used as a substitute for magnitude) relations for Classical Cepheids, type II cepheids and RR Lyrae. Figure 7 from Soszyński et al. (2011b).</td>
</tr>
<tr>
<td>1.9</td>
<td>Original plot from Leavitt &amp; Pickering (1912). Apparent Magnitude vs log(period) for Cepheid variables in the Small Magellanic Cloud. The two distributions correspond to the maxima and minima of the objects.</td>
</tr>
<tr>
<td>1.10</td>
<td>Slope (and associated errors) of period-luminosity relations of LMC cepheids. The green points are theoretical model and yellow are the observed data. As can be seen, at longer wavelengths the relation has less scatter. Figure 1 from Subramanian et al. (2017).</td>
</tr>
<tr>
<td>1.11</td>
<td>Period-Luminosity relation for a sample of Gaia DR2 cepheids in the V Band. Part of Figure 4 from Lazovik et al. (2019).</td>
</tr>
<tr>
<td>1.12</td>
<td>Period-Luminosity relation for a sample of Cepheids using the four WISE bands. Black dots are fundamental mode classical Cepheids and red crossed are first-overtone classical Cepheids. Figure 2 from Wang et al. (2018).</td>
</tr>
</tbody>
</table>
1.13  **Period-Luminosity relation for variable red giants in the LMC. RGB**
OSARGs are light blue, AGB OSARGs are dark blue, Miras and SRV are pink if O-rich and red if C-rich. The green points are stars with long secondary periods. Figure 1 from Soszyński et al. (2007)....

1.14  **Sample light curves for different types of eclipsing binaries from VISTA.** The spotted type refers to eclipsing variables such as RS Canum Venaticorum, where the presence of sun spots can have an additional noticeable effect. Figure 13 from Angeloni et al. (2014)....

1.15  **I band model of overlapping variability during eclipse of a Cepheid.**
Bottom plot is light curve and model, upper plot is residuals of the model. Object measured by OGLE (Pilecki et al., 2013). ................

1.16  **Color coded HR diagram using variables from the Hipparcos catalogue.** Figure 4 from Eyer & Mowlavi (2008)......................... 26

1.17  **Evolutionary tracks of pre-main sequence stars, labelled by mass.** The Hiyashi track is visible as the vertical paths undergone by lower mass stars. Image taken from Figure 2 of Palla (2012) ..................

2.2  **Comparing the amplitudes and light curves of LPVs compared to eclipsing binaries.** The image shows phased lightcurves for an eclipsing binary (left) and an LPV (right). The two horizontal lines are magnitude limits used by the authors to aid in picking out eclipsing binaries. Figure A.1 from Groenewegen (2005) ............ 40

2.3  **Examples of OSARG I-band phased light curves (top 4 images) compared to phased light curves for LPVs and ellipsoidal variables with red giant components (below).** Top 5 images sourced from OGLE team (Soszyński et al., 2013a). Lower image sourced from Figure 1 in Soszyński (2022) .................................................... 42

2.4  **Distribution of stars in the Bayestar19 (Green et al., 2019) extinction maps, plotted in solar-centric Cartesian coordinates.** The black dot denotes the galactic centre. The colour scale from blue to yellow is an arbitrarily assigned logarithmic scale of density. Figure 17 from Green et al. (2019) ................................................................. 46

2.5  **Dust density in the Lallement et al. (2019) extinction maps along the plane with the sun at the centre and the galactic centre to the right.** The colour scale shows the extinction strengths. The dotted line is the limits which are the best quality resolution of 25 pc grid size can be achieved. Figure 9 from Lallement et al. (2019) ............ 47
2.6 Dust density at different vertical planes off the galactic plane. Colour scale and black line are shared from 2.5. Figure 10 from Lallement et al. (2019).

2.7 Dust density at different vertical planes off the galactic plane. Colour scale and black line are shared from 2.5. Figure 11 from Lallement et al. (2019).

2.8 Extinction in the galactic plane for \(|b| < 0.4\). The y-axis represents the galactic longitude, the position of the sun is marked with an X and the galactic centre with a +. The lower graph is the results from a previous 3D extinction map by Marshall et al. (2006). Figure 16 from Chen et al. (2013).

2.9 Compilation of IR extinction laws from literature. Red stars plot the extinction law towards the galactic centre derived by Fritz et al. (2011). Blue triangles are derived from stars towards the galactic centre by Rieke & Lebofsky (1985). Green squares were derived from red clump giants by Nishiyama et al. (2009). Cyan diamonds are galactic center extinction found by Gao et al. (2009). The remaining three symbols are extinction laws obtained via sight-lines not towards the galactic centre. The solid and dash-dot lines denote the model extinction curves - with \(R_V = 3.1\) and \(R_V = 5.5\) respectively. Figure 1 from Gao et al. (2013).

2.10 An example of the VVV single epoch detection limits, as they are used to remove spurious detections and rejected sources selected in the mag-err plane. The darker line is fit to the logarithm of the errors at a given magnitude, whereas the grey line is the sigma-clipping limit. The darker objects are the rejected objects and the grey ones are the retained sources. Figure 2 from Mauro et al. (2013).

2.11 Sample distributions of magnitude and colour, as used in NWAY’s Bayesian analysis process. In the top example, the histograms of WISE objects with counterparts in the VIVA catalogue (selected) versus the WISE objects with counterparts in the complete VVV catalogue (all). This creates two distributions, one corresponding to variable candidates and one corresponding to both variable and non-variable objects. In the bottom example, the \(K_s - 3.6\) colour - calculated using the VVV’s \(K_s\) band and the glimpse 3.6 micron band is used to do the same. These distributions are used by NWAY to probabilistically weight all candidates of matching sets of objects.

2.12 Mean-Absolute Deviation (MAD) against median magnitudes for a sample of VVV objects in clusters. Figure 1 from Alonso-García et al. (2021).
2.13 (Top) Light curve of RR Lyrae variable star, phase folded, sourced from the WFCAM [Casali et al., 2007] calibration archive. Object is part of the VIVA crossmatched catalogue. Additional points on the light curve using data from Pan-STARRS (points labelled ip1, zp1 and so on), SDSS (points labelled r, g and so on) and NEOWISE (points labeled W1, W2) are coloured according to the label on the right. (Middle) Fluxes minus the average value in each band. The black point labelled roots denote points of interest such as maxima and minima. As can be seen from this and the above figure, the light curve shows the rapid peak and slow decline characteristic of a RR Lyrae, although with some scatter. The scatter is potentially due to the Blazhko effect [Szabó, 2014], which results in variation of period and amplitude over time - since this data has been collected over several years. (Bottom) The broadband SED of the above data. The spectrum peaks at around ~ 3000K, which does not fit the expect temperature [Catelan & Smith, 2015b].

2.14 The behavior of effective temperature calculated using IRFM as a function of (BP - RP) colour for dwarf and giant stars (top and bottom respectively). The stars are grouped according to metallicity, and the different polynomials (corresponding to each metallicity) are drawn using the same grouping. Figure 1 from Mucciarelli & Bellazzini [2020].

2.15 The SED for VVV object id 515838123764 after adjusting for low extinction, the objects extinction in our extinction map and under high extinction.

2.16 Using Gaia temperatures for comparison, the performance of the IRFM method on a random sample of stars. Both axis display the temperature in Kelvin and the black line denotes a 1 to 1 match. The scatter graph is coded based on density of points in that region.

2.17 Crossmatched periods - labeled literature periods here - against the periods calculated using the 5 different methods used by VIVA. The periods are plotted in log10 base, and the scatter plot is coloured according to its density. The objects used here are the subsample VVV-CVSC-CROS - defined as objects with matches in the AAVSO or SIMBAD database.

2.18 Crossmatched periods - labeled literature periods here - against the 4 highest periodogram spikes for the Lomb-Scargle algorithm. The first graph corresponds to the highest periodogram spike, the second corresponds to the second highest periodogram spike and so on. The periods are plotted in log10 base, and the scatter plot is coloured according to its density.
2.19 Crossmatched periods - labeled literature periods here - against the highest periodogram spikes for the Quadratic Mutual Information minimization algorithm. The first graph corresponds to the highest periodogram spike, the second corresponds to the second highest periodogram spike and so on. The periods are plotted in log10 base, and the scatter plot is coloured according to its density. 78

2.20 Mean magnitude across the matched Ogle OSARGs and their VVV counterparts. KsAperMag3 denotes the VVV Ks band magnitude with a 2.0 arcsecond aperture. 79

2.21 Mean amplitude of OGLE Bulge Mira and VVV counterparts. The vertical axis denotes the relative density in each histogram bin. 81

2.22 Mean amplitude of OGLE Bulge Cepheids and VVV counterparts. The vertical axis denotes the relative density in each histogram bin. 82

2.23 Light curve of OSARG, showing both VVV and OSARG data for a matched object. Object ID in VVV DR5: 51581365001 83

2.24 Light curve of OSARG, showing both VVV and OSARG data for a matched object. Object ID in VVV DR5: 51583932380 84

2.25 Light curve of OSARG, showing both VVV and OSARG data for a matched object. Object ID in VVV DR5: 51584725437 85

2.26 Mean amplitude of OGLE OSARGs and VVV counterparts. The y axis describes the density in each histogram bin and the x axis shows the amplitude in each bin for the VVV Ks band and the OGLE I band. VVV objects with magnitudes approaching saturation threshold (<13 mag) were removed, using a total of 147585 objects. 86

2.27 Mean amplitude of OGLE Bulge Mira and VVV counterparts. The vertical axis denotes the relative density in each histogram bin. 87

2.28 Mean amplitude of OGLE Bulge Cepheids and VVV counterparts. The vertical axis denotes the relative density in each histogram bin. 88

2.29 VVV tilestack for object ID in VVV DR5: 515910962364. The bright object is visible in the centre whilst a saturated object with an empty spot in the centre is visible above it. 89

2.30 Filter efficiency curves for the filters used in this project. The plots take the form of filter transmission against wavelength. Filter data provided by the SVO filter profile service (Rodrigo et al., 2012; Rodrigo & Solano, 2020) 93
2.31 Light curves of a sample OSARG using panSTARRS data. The additional points present in the $z$ and $y$ filters are the mean values of the VVV data in these bands, in the VVV counterparts of those filters. Object ID in VVV DR5: 515909698212

2.32 Atmospheric parameters of the 769 stars in the X-Shooter library. Star shapes denote slightly older values by Gonneau et al. (2017) and circles are more recent values from the first data release by Arentsen et al. (2019).

2.33 Two X-shooter models normalized to the higher amplitude spectra, with the filters overlaid for comparison of wavelength position. The orange coloured object is X-Shooter object 374, HD 134439 and the blue coloured object is X-Shooter object 300, HD 170756. HD 170756 is believed to be a post-AGB star with temperature around $\sim 5300$ K and HD 134439 is believed to be a subgiant star with temperature around $\sim 5100$ K. Both have similar metallicity. The noticeable “blurry” regions at 17000 Å and 13000 Å are examples of regions affected by Slit Loss Correction and are the two biggest regions affected. There is also a similar area around 5000 Å, which overlaps the OGLE V band filter.

3.1 Comparison of the IR spectrum of 2MASS suspected Mira J1835 (black) with a red supergiant (orange) and a late-type Mira variable (red). Regions of low atmospheric transmission are shown in grey. Figure 2 from De et al. (2022).

3.2 Samples of 1000 Miras, SRVs and OSARGs (Top, Middle, Bottom) plotted according to their VVV apparent magnitude against their Bailer-Jones et al. (2021) distances using photometric information. The error for each point, both in terms of magnitude and the upper and lower distances predicted are shown.

3.3 Samples of 2000 Miras, SRVs and OSARGs (Top, Middle, Bottom) plotted according to their VVV apparent magnitude against their Bailer-Jones et al. (2021) distances, this time calculated only using geometric information. The error for each point, both in terms of magnitude and the upper and lower distances predicted are shown.

3.4 Photometric (Top) vs Non-photometric distances (Bottom) (Bailer-Jones et al. 2021) compared for samples of 2000 Miras, SRVs and OSARGs. The height of any point describes the fraction of each object contained at that distance.
3.5 Colour (J-Ks) vs. Ks Magnitude (Corrected Mag, adjusted for extinction) diagram for OGLE-sourced Miras in VVV, coloured by number of objects. Extinction for both J and Ks is corrected using the extinction maps of Schultheis et al. (2014) using the objects position on the sky and distance from the Bailer-Jones et al. (2021) parallax distances. ................................................................. 110

3.6 Colour-Ks Magnitude diagram for Miras. The circles and empty squares denote C-rich and O-rich Miras respectively. The pluses and stars are objects using 2MASS photometry. The grey dots and crosses denote O-rich and C-rich OGLE LMC Miras respectively, which have a correction of 4 magnitudes to simulate their potential bulge magnitudes. Figure 6 from Matsunaga et al. (2017) ............... 111

3.7 Colour (J-Ks) vs. Ks Absolute Magnitude diagram for OGLE-sourced Miras in VVV, coloured by number of objects. Extinction correction is performed as described in Figure 3.5................................................................. 112

3.8 The absolute K magnitude against log period for a sample of Miras in dwarf spheroidals in NGC 6822. The points are colour coded as Fornax (cyan), Leo I (magenta), Sculptor (green) and Phoenix (black). The black trend line is the PL relation for the LMC. Figure 6 from Whitelock (2012).......................................................... 113

3.9 The absolute K magnitude against log period for a sample of local Miras (squares) and SRVs (triangles). The labelled diagonal regions (B+, C’ and C) denote reproductions of the period-magnitude sequences of Ita et al. (2004). B+ describes the upper part of the sequence for SRV stars and C’ and C denote Mira variables. Figure 9 from Sudou et al. (2019). ...................................................... 114

3.10 Colour (J-Ks) vs. Ks Magnitude (Corrected Mag, adjusted for extinction) diagram for OGLE-sourced SRVs in VVV, coloured by number of objects. Extinction correction is performed as described in Figure 3.5................................................................. 116

3.11 Colour (J-Ks) vs. Ks Absolute Magnitude diagram for OGLE-sourced SRVs in VVV, coloured by number of objects. Extinction correction is performed as described in Figure 3.5................................................................. 117

3.12 Colour (J-Ks) vs. Ks Magnitude (Corrected Mag, adjusted for extinction) diagram for OGLE-sourced OSARGs in VVV, coloured by number of objects. Extinction correction is performed as described in Figure 3.5................................................................. 118

xviii
3.13 Colour (J-Ks) vs. Ks Absolute Magnitude diagram for OGLE-sourced OSARGs in VVV, coloured by number of objects. Extinction correction is performed as described in Figure 3.5. ........................................ 119

3.14 From a sample of 1000 SRVs originally detected using OGLE, with counterparts in VVV and Gaia EDR3 distances using Bailer-Jones et al. (2021). Plot 1 denotes objects plotted according to their galactic latitude by their distance. Plot 2 denotes objects plotted according to their altitude above the galactic plane with respect to their distance. The colour of each point denotes the galactic longitude, relegated to a third axis since I wanted to better understand what inclinations the ∼ 7 kpc predicted objects (Bailer-Jones et al., 2021) lay at. ........................................................ 121

3.15 The impact of each indices from the crossmatched VIVA-Gaia dataset on distinguishing between the various categories. The colour scale denotes the amount of positive correlation (yellow colours) and anti-correlation (blue colours). .......................................................... 133

3.16 A two component t-SNE clustering with perplexity 30. .................. 134

3.17 A two component t-SNE clustering with perplexity 40. .................. 135

3.18 A graphical example of the difference between the nodes of a random forest tree (upper) and those of a PRF (lower), showing the process for an object propagating through the node. The random forest makes a simple boolean decision with one of two outcomes. The PRF sends the down both paths with a different weighting on each - based on the uncertainty of the feature used for that node. Figure 1 from Reis et al. (2019) ......................................................... 137

3.19 Upper, a simple random forest. Lower, a simple PRF. A comparison of the path down an individual decision tree an object takes for each algorithm. In a normal random forest, an object follows a single trajectory down the orange line of nodes and arrives at the final red node outcome. In a PRF, each node propagates it’s input (as well as the probabilities) to each child node, ending up with a prediction at each end node. ............................................................ 138

3.20 A trimmed PRF. If the output to a node would fall below an assigned threshold, we can drop that node, removing it from further calculation. This is used to trim nodes where the probability of it being the right one is low enough to remove it, aiding the speed of the network by removing parts it has to calculate. Black nodes and paths denote trimmed pathways. ................................................ 139
3.21 Confusion matrix of the first section of the PRF - denoting the fraction of each object, by true label, and where they were assigned by the PRF. A value of 0.91 would correspond to 91% of all objects in that row assigned to the label corresponding to that column. In the example here, 12023 of the eclipsing variables were labelled as such, whilst 1215 were labelled incorrectly. The following abbreviations are used: ecl - Eclipsing Binary, pul - Pulsating variable, yso - Young Stellar Object.

3.22 Confusion matrix of the second section of the PRF

3.23 The difference in statistical variance and distribution of Skew (x-axis), one of the common derived features following the work of Nun et al. (2015). As can be seen there are noticeable differences in common features between surveys. Figure 2 from Aguirre et al. (2019).

3.24 A step by step example of the convolutional process, demonstrating sliding the window or kernel across the data. The size of the kernel and the step size are hyperparameters, set before training. The values used will be discussed in section 3.4.2. Figure 5 from Aguirre et al. (2019).

3.25 Confusion matrix of the CNN - denoting the amount of each object, by true label, and where they were assigned by the CNN. The following abbreviations are used (in addition to SRV, Mira and OSARG which have already been defined): Cep - Cepheid, RRlyr - RR Lyrae, ContactBin - Contact Eclipsing Binary, DetachedBin - Non-Contact Eclipsing Binary.

4.1 The Ks band magnitudes violin plot for the two sets of predictions. Showing the results of the CNN (top) and the PRF (bottom). The graph denotes the distribution of the Ks magnitude along the vertical axis.

4.2 The Ks band amplitudes violin plot for the two sets of predictions, showing the variational amplitudes of the objects in each category. Showing the results of the CNN (top) and the PRF (bottom).

4.3 Comparing the predictions of the CNN and PRF, showing how each network classified each object and the amount of overlap or disagreement in their predictions. The most notable disagreement in the predictions is the PRF classifying ~ 50% of the CNN-predicted LPVs as eclipsing binaries - outnumbering the number of eclipsing binaries both networks agree on.
4.4 The overlapping Ks magnitude histograms for all VIVA objects, VIVA objects with Gaia counterparts and VIVACE. The vertical axis denotes the fraction of objects at each magnitude due to imbalances between the 3 catalogues.

4.5 The Period-Absolute Ks band magnitude plot for the sample of LPVs found by the CNN, as described above. The red line shows the location of the Sun et al. (2022) period-luminosity relation for Mira, as described in Section 4.3. The period here is found using the Lomb-Scargle periodogram.

4.6 As Figure 4.5, but zoomed in on the seemingly structured regions of period - the vertical lines.

4.7 The Period-Wesenheit relation for the original sample of literature-crossmatched OSARG stars in VIVA. A logarithmic axis has been used to include the several high period objects.

4.8 The Period-Absolute K band magnitude relation for the sample of local variables from the Bulge Asymmetries and Dynamical Evolution (BAaDE) project (Quiroga-Nuñez et al., 2020). The absolute K band magnitudes are calculated using distances acquired from Gaia parallaxes. The literature LPV sequences from Whitelock et al. (2008) and Matsunaga & IRSF/SIRIUS Team (2007) are solid and dashed lines respectively and the sequences with letters corresponding to different pulsation modes (Lebzelter et al., 2019). Modes C and C’ are fundamental and first overtone Mira and mode D denotes long secondary period objects. Figure 13 from Quiroga-Nuñez et al. (2020).
# List of Tables

3.1 A tallying of the variability indices used in modern astrophysics studies and reviews. This is not a complete sample of every index used and some have been combined with others for the sake of brevity. The standard deviation \( \sigma \) is used by itself as well in more complex indices such as \( \frac{\sigma}{\mu} \). The von Neumann ratio is often used as \( \eta \). The Stetson indices have clipped and time weighted versions described in their section that are merged with the base version. The references are aliased as follows: 2: Shin et al. (2009) 3: Sokolovsky et al. (2017) 4: Pashchenko et al. (2018) 5: Sokolovsky & Lebedev (2018) 6: D’Isanto et al. (2016)

3.2 The precision, recall and f1 corresponding to the above PRF, for each object type and overall.

3.3 The precision, recall and f1 corresponding to the CNN, for each object type and overall.

4.1 CNN Predictions of final sample of candidate variables. Results are separated into the broader superclasses and variable subcategories. Eclipsing includes Contact and Detached Binaries, LPV includes OSARGs, SRVs and Miras and Other Pulsating includes Cepheids and RR Lyrae.

4.2 PRF Predictions of final sample of candidate variables.

4.3 A comparison of Slope (a) and Zero-point (b) for a collection of Miras in the local group.
Chapter 1

Background

Variable stars are one of the most important discoveries in Stellar Astrophysics. The term variable describes behavior in the middle ground, between the macro-scale changes as part of the slow evolution over millions to billions of years that all stars show and the micro-scale atmospheric changes over minutes that a star like Sol exhibits. This middle ground covers changes in luminosity on timescales (Catelan & Smith, 2015b) that are observable to us, and detectable even from our distance to them.

The variations observable to us depend on which object and instrument - observing the micro-scale variations over seconds in White Dwarfs to the decade long variations observed in some evolved stars. As more and more observations are gathered we may reach a point where century to millennia long variations (Templeton et al., 2005) will be commonly observed. This ties into the choice of instrument - the GALEX space telescope can use its 30 minute exposures to measure activity in 15 second windows (Rowan et al., 2019), ideal for detecting White Dwarfs with periods of less than 1 day but impractical for longer period objects or ones where much of the variability lies in the optical or infrared bands.

The means by which these stars vary, and the properties of the variations are as varied as stars themselves can be. There is also a lot of overlap in periodicities and timescales between different categories of variables, even though the actual range of periodicities and timescales of variables is large. Categories are created based on similar light-curve shapes, or when underlying physical properties are understood (Samus’ et al., 2017). Some variables display increases of brightness of the same amplitude each time in intervals with a fixed period, whereas some
variables erupt in large outbursts with no predictability.

These stars have shown to have great scientific importance to Astrophysics in general, from Hubble showing the Andromeda Galaxy was a separate object akin to our own galaxy [[Hubble 1929]] using Cepheid variable stars to estimate its distance, to being some of the earliest steps in the cosmic “distance ladder”, a collection of methods which can determine the distance to cosmic objects [[Freedman et al. 1997]]. Each "rung" in the ladder builds on the ones before, which are used to help calibrate and find correlations between the earlier steps.

The distance ladder begins with parallax measurements for measuring distances to local objects. With this important step in place, we can begin to calibrate distance relations to standard candles - objects with predictable brightness. The period-luminosity relations of variable stars such as Cepheids or RR Lyrae are calibrated using these known distances, giving reliable means to convert simple observable information such as the frequency of a Cepheids variations to absolute magnitudes to distances [[Bohm-Vitense 1997; Riess et al. 2016]]. Now calibrated, these relations can be applied outside of the distances parallax can be measured. Bright pulsating stars such as Cepheids can be used to continue this process - by using Cepheids with known distances in galaxies with Type Ia supernova, this absolute magnitude and distance of these objects can be calibrated as well, adding another rung to the distance ladder. The last main step on the distance ladder is redshift, calibrated using Type Ia supernova and giving a well established starting point into cosmological scales [[Hogg 1999]].

Looking closer to home, the distance measurements from variable stars can aide the study of our own galaxy. Even after many years of study, the distribution of stars and structure throughout the galaxy is still a topic of active research to this day [[Johnson et al. 1961; Alonso-García 2021; Griv et al. 2021]]. In particular, the regions around and behind the Galactic Plane are very obscured by dust from our perspective which can obscure many objects and bring them below the detection limits of observations, particularly in optical or ultraviolet [[Fitzpatrick 1999]].

In addition to distance variable stars are important in other ways. Microlensing, the phenomenon where an eclipsing foreground object can act as a lens for a background object, can be greatly affected by any variations in the lensing star [[Sajadian & Ignace 2020; Alcock et al. 2000]]. Common practice would be to remove any microlensed object with a lens showing a binary companion
or other small scale variation, but some research has shown that with proper understanding of the variable source this could be mitigated or even turned to an advantageous tool (Assef et al., 2006). Microlensing share some similarities with extrinsic variable stars, defined by changes in brightness caused by a change in the amount of light that reaches our detectors and is not grouped along with variable stars seemingly by tradition. Another field variable stars can help is in understanding the variability behind young stellar objects or novae-like sources and thus giving us ways to study the disks of material around transient objects or in star forming regions (Fernandes et al., 2018), or likewise as as probes of astroseismology - using stellar oscillations to infer the physics of stellar structure (Yu et al., 2020).

Over time the list of variable sources has increased dramatically, from ancient astronomers catalogues of one or two very noticeable candidates to the recent surveys of the past two decades. As the new millennia started there were surveys such as IRAS, MACHO and OGLE and as technology has developed there are more large scale surveys of the sky with higher spatial resolution and greater depths, such as Pan-STARRS, ZTF, ASAS-SN, WISE and Gaia and KMT-Net. These newer large-scale sky surveys are rapidly becoming some of the most important data sources in modern astrophysics, capable of observing millions or billions of stars and finding 10000s to 100000s of variable stars (Heinze et al., 2018; Clementini et al., 2018). As these surveys become larger and more powerful, they make great strides in aspects of observing in order to cover new ground and push the boundaries of what we can observe. Newer surveys feature advances like wider fields, observations at multiple wavelengths, observations over multiple epochs, spectroscopic data and astrometric data such as the recent Gaia survey (Gaia Collaboration et al., 2018a).

The vast numbers of objects across unprecedented scales push the limits of what we can study but with newer data science based-techniques, we stand to learn much about individual objects and populations or groups with common properties.

With the ability to get a better view of the variability of our galaxy, it is important to understand what we can learn from each survey source. The surveys mentioned above cover different areas of the sky and wavelengths and observe in different patterns with different instrumental quirks.
1.1 Sources Used and Other Surveys

The data from the VISTA Variables in the Via Lactea (VVV) \citep{Minniti10} survey are the primary resource used in this project. VVV is a public survey on the Visible and Infrared Survey Telescope for Astronomy (VISTA) using the VIRCAM near-infrared camera, covering 570.3 sq. deg. \citep{Minniti10} of the Galactic Plane (galactic longitude between -65° and -10° and galactic latitude between -2° and +2°) and Bulge (galactic longitude between -10° and +10° and galactic latitude between -10° and +5°).

VVV covers much of the bulge and disk of the Milky Way, areas avoided by most optical surveys. The heavy extinction and crowding limit the depth for visual band instruments. A survey like VVV is more suitable for the mission of observing deeper towards the galactic centre.

VVV is observed in 5 broadband filters ($Z, Y, J, H, K_s$), with filter coverage shown in Figure 1.1 and each pointing was observed in $\lesssim 10$ epochs for the $Z, Y, J$ and $H$ filters and over 100 $K_s$ epochs for the bulge and 60 $K_s$ epochs for the disk over its 5 year campaign. The VVV Data Release 4 (DR4) \footnote{https://www.eso.org/sci/publications/announcements/sciann17009.html} contains $10^9$ unique sources, of which 1-10 million are likely to be variable stars \citep{Minniti10} with VVVDR5 containing even more epochs of observation on these objects. VVV is designed to be capable of finding common variable stars like RR Lyrae even through 10 magnitudes of visual extinction \citep{Minniti10}.

It is the first wide-sky map in the near-IR to have multiple epochs of data, totalling 1929 hours. The Near-IR camera covers a 1.65 degree diameter field of view made up by 67 million pixels of mean size 0.339 arcseconds. The field of view of the VISTA instrument is displayed in Figure 1.2.

What makes the VVV survey interesting with respect to other large scale variability surveys is its coverage in the near-IR, which comes with many advantages such as being able to probe deeper into the reddened, dusty center regions of the galaxy. Figure 1.4 shows the large difference in $K_s$ band extinction limits to detect a Cepheid variable with a period of 10 days, showing that modern optical surveys such as Gaia can detect their test variable up to $\sim 6.5$ kpc, whereas in near-IR the same variable can be observed at 10 kpc, with strong extinction in both cases \citep{Matsumaga17}. Figure 1.3 similarly demonstrates the extinction...
Figure 1.1  *Filter efficiency of the VVV filters, with the quantum efficiency of a sample detector plotter in pink for comparison. Filter properties image from the Cambridge Astronomy Survey Unit website.*

an optical survey must face when observing in very dusty regions.

The use of near-IR affects what variables we can observe. Hot stars are difficult to distinguish from cooler stars using near-IR colours alone, but using optical and near-IR makes them distinguishable. However, near-IR selection will make hot stars rarer than an optical selection. Objects that pulsate strongly in the optical and especially UV, but with little change in the IR will be harder to detect. For example Blue-White pulsating variables, which emit mainly in the higher visible part of the spectrum due to their high temperatures will be less frequently observed as variable with the near-IR VIRCAM (VISTA InfraRed CAMera) on the VISTA telescope. High extinction and reddening due to dust in the Galactic Center regions adds further complications to this - limiting the optical data we can gather and reddening blue objects such that they can be confused with redder unaffected stars.
Another factor that affects what types of variable object can be detected is the timing and number of observations. VVV has observed a total of 100 epochs for the bulge and 60 epochs for the disk (Catelan et al., 2013), with light curves being measured in batches of 2-6 with each observation on the order of a minute each (Dalton et al., 2006) over a single night and intervals of typically a hundred to three hundred days between consecutive observations (Catelan et al., 2013). These batches of 6 overlapping observations - pawprints - are used to create a tile. VVV observes in multiple bands, but not equally. Each object has between one and two observations in the Z, Y, H and J bands, with the $K_s$ band typically having 50-100. An extension of VVV, VVVX (Minniti, 2016, 2018) has ran with the aim of covering a wider area of the sky but with fewer epochs, including 3–5 epochs of observation in the original survey area for monitoring long term variables. The coverage of the larger area does not provide enough epochs for practical variability identification based on light curves (see Section 2.3.2).

Although it is possible to observe variable objects with periods less than a day using time-sparse data like this some variables, such as pulsating white dwarf stars, typically pulsate with periods of minutes (Catelan & Smith, 2015c), resulting in difficulty deducing periods or even identifying such objects as variable. The effect of having too slow a sampling rate to properly capture the signal - a regular time varying process, in this case the stars pulsation, is known as
A comparison of the decrease in magnitude one would observe through equivalent amounts of extinction between optical and near-IR. The dotted and dashed lines denote different extinction laws; a simple power law for the filled lines and Cardelli et al. (1989)'s law of extinction for the dashed lines. The grey and black colours are for Gaia G band and V band respectively. Figure 2 from Matsunaga (2017).

aliasing. Trying to analyze the aliased signal will lead to bad measurements for period and amplitude (Baluev, 2012). Hot Subdwarf Pulsators have a similarly short period (Catelan & Smith, 2015d). Although both types can be relatively luminous (as seen in Figure 1.5), they pulsate with very low amplitudes (Catelan & Smith, 2015c,d) and are relatively hot and thus harder to see in the near-IR. For these reasons it was decided that these categories will not be the focus of our observations.

As mentioned in Section 1.2.1 certain types of variable can be more easily identified in the near-IR instead of the optical because of a narrowing of the instability strip in colour-magnitude space in the IR compared to the optical causing a tightening around Period-Luminosity relation (Subramanian et al., 2017). This is leads to the VVV survey being well suited to the goals of this project and finding some of the most scientifically useful variable stars such
Figure 1.4 The maximum extinction at which a 10 day period cepheid can be detected with a given limited magnitude. The red area and dotted line show a real example case, with extinction (and error bar) increasing with respect to depth, drawn from a 3D extinction map, whereas the dot-dash line indicates the extinction increasing at a constant rate. The dotted and dashed lines denote different extinction laws; a power law similar to Nishiyama et al. (2006)'s law of extinction for the filled lines and Cardelli et al. (1989)'s law of extinction for the dashed lines. Figure 3 from Matsunaga (2017).

as Cepheids (see Section 1.2.1). Another type of variable that the survey is well suited for are the Long Period Variables (LPVs). Their red colours make them excellent targets for observations in the near-IR, and their large magnitude changes are quite noticeable. In addition, they are quite common with over 200,000 having been detected toward the Galactic bulge in the OGLE-III catalogues (Soszyński et al., 2013a) due to the majority of stars on the Red Giant Branch (RGB) or Asymptotic Giant Branch (AGB) showing instability or variability in some form (Soszyński et al., 2013a).

For the purposes of variability studies like this project, there are challenges associated with operating in this wavelength as well. The relative newness of large near-IR surveys also means that many variable stars have been studied relatively
Figure 1.5  A figure of the evolutionary position of the pulsating variable stars on the HR diagram. Distribution of different types of pulsating variable stars across the HR diagram. The solid black line is the zero-age main sequence (ZAMS). The dotted lines labelled with numbers represent evolutionary tracks for a star of that mass. The dot and dash line is the zero-age horizontal branch. The dotted lines to the left are the white dwarf cooling curve for He and CO white dwarfs. The different shading patterns represent the different pulsation mechanisms - diagonally to the left are $p$-mode, diagonally to the right are $g$-mode pulsations. Vertical lines are non-adiabatic pulsations, horizontal lines are solar-like pulsations. Figure originally from Jeffery (2008), recreated by Catelan & Smith (2015a).
little at longer wavelengths (Angeloni et al., 2014). As will be expanded upon in Chapter 3, having a large number of variable stars covering a wide range of types allows the use of them as references to identify new variables by comparing them to old ones. Angeloni et al. (2014) introduces the VVV template project, which is a database of high quality near-IR light curves for a variety of variable stars. This is still an ongoing project and the numbers of variables used are still significantly smaller than similar databases for visible light curves (Angeloni et al., 2014; Heinze et al., 2018).

One possible way to expand our effective light curve database will be to combine VVV’s near-IR data with optical data from other surveys, which will allow the use of visual light curves for variables from sources such as the Optical Gravitational Lensing Experiment III (OGLE-III) (Szymański et al., 2011) and the Asteroid Terrestrial-impact Last Alert System (ATLAS) data (Heinze et al., 2018). By cross-matching between this catalogue and others including the Near-Earth Object Wide-field Infrared Survey Explorer (NEOWISE) catalogue (Mainzer et al., 2014), Gaia DR2/3 (Gaia Collaboration et al., 2021) and the Panoramic Survey Telescope and Rapid Response System (PanSTARRS) catalogue (Chambers et al., 2016), the variable stars present in the survey can be identified (using a variety of indices measuring the scatter and spread of the collected data), examined (classifying the found objects by type where possible and analyzing them to determine stellar parameters such as surface temperature), and distance information obtained (from how extinction and distance have made the star appear to be at different brightness and temperature than it actually is), which help map the innermost regions of the Milky Way and the far side of the Galaxy.

Another advantage of cross-matching is a much broader waveband coverage for objects that can be found to have matches across the different sources. The rational for needing broad waveband coverage like this is described in detail in Section 2. The Pan-STARRS 3π sky survey (Chambers et al., 2016) and SDSS (York et al., 2000) extend our broadband data into the visual bands, whereas WISE (Wright et al., 2010) and NEOWISE (Mainzer et al., 2014) extend into the mid-IR. The NEOWISE public data release includes multiple epochs of data, and this mid-IR coverage works well with the dusty regions VVV is observing and provides data about the cooler ends of the spectrum of many stars - important since many long period variables stars such as Miras are relatively cool.

Instead of initially testing the analysis pipelines created in this project on the
VVV dataset, I have been using the Wide Field Camera (WFCAM) on UKIRTs calibration data (WFCAMCAL) as a test set. WFCAM uses very similar filters to VVV, and this dataset has similar number of epochs in the K band as VVV does in its $K_s$ band. This dataset has the advantages of existing classification and identification of variables (Ferreira Lopes et al., 2015). It also has less issues with high extinction and crowding of sources due to covering a wide range of fields over the northern hemisphere, covering regions of the sky much less densely packed than the Galactic Centre. Towards the end of my PhD an expanded sample of identified variables was released - VIVA (Ferreira Lopes et al., 2020) which helped to expand upon the number of test cases and will be discussed in more detail later.

In Section 1.2 I will provide a summary of the types of variable stars that has been discovered so far, identifiable characteristics and their relevance to the project because of the scientific information they provide. In Section 1.1 I will discuss how advancements in large scale surveys in recent times has furthered the development in this field. In Section 2 I will discuss various algorithms and techniques for identifying variable stars from large datasets and automatically identifying and extracting useful information such as variable type from them.

### 1.2 Variable Stars

The subject of variable stars is broad and for ease of understanding and clarity, the topic will be separated into several sub-groups and their relevance to this project will be discussed separately.

One can separate all variables stars into two broad groups, those that show intrinsic variations and those that show extrinsic variations. Extrinsic variations are caused by some factor changing how much light from the star reaches us, either from obscuration by another body or phenomenon in the stars rotation. Intrinsic variations occur from some change to the star itself, including the star expanding and contracting due to internal processes or outbursts caused mass loss or accretion. See Figure 1.6 for an overview on the many types of variable objects that can be observed and their many subgroups. Most of these will be covered in Sections 1.2.1 to 1.2.6, these sections being ordered by relevance to this thesis, an explanation of each contained in each section.
1.2.1 Intrinsic: Pulsating

Pulsating variable stars are some of the most well known examples of variable star, dating back to the 1600s with the discover of the variable Omicron Ceti (Wallis, 1693), better known as Mira. Later that century, the hypergiant P Cygni was observed to brighten and fade out of sight, only to reappear years later. Most of types of pulsating variable are displayed at their positions on the HR diagram in Figure 1.5 while each variable in this category can also be divided into sub-categories which are described below:

The Cepheid-like variables (or Instability Strip variables, see region contained by the dashed line in Figure 1.5) are similar to the prototype star Delta Cephei, the first Cepheid variable star discovered. This grouping encapsulates a variety of stars which have proven scientific value over the years. It includes the Delta Cephei (Type I Cepheids or Classical Cepheids), Type II Cepheids, BL Boo stars (Anomalous Cepheids), RR Lyrae, Delta Scuti, SX Phoenicis, Rapidly Oscillating Ap (ROap) and Gamma Doradus stars. All of these types occupy a similar region of the HR diagram, known as the instability strip. Each type corresponds to a
certain region of the HR diagram, and stars of different temperature, mass and age.

The mechanics behind the instability strip and most other pulsating variables derives from the interplay of opposing disturbing and restoring pressures inside the star [Sandage & Tammann 2006]. In contrast to much more energetic instabilities that runaway and outstrip the natural damping effect the star would have on itself, eventually leading to drastic changes in the star, such pulsations lie in the sweet spot where the oscillation exceeds the stars damping, but saturates and caps off before it can become unstoppable and continue to grow. The star is unstable on a micro scale, but stable on the macro scale. It then remains in that state until something, such a change in the stars burning, causes the balance to tip one way or another.

The primary mechanism in these circumstances is the Kappa or k mechanism [Deng & Xiong 2001; Marconi & Palla 1998]. Below the stars photosphere the temperature increases and allows He II to form in a layer. As this layer contracts, it rises in density and temperature and the ionization to He III begins. This dense layer of He III is largely opaque to further ionization and the stars rising internal temperature and radiation pressure outscales the restoring force of pressure or gravity and forces the layers of the star outwards [Gautschy & Saio 1996]. This increase in size of the star is what causes the increase in magnitude we observe - the star has more surface area to emit from. As the outer layers expand, the He III begins to recombine to He II, the opacity falls, the radiation pressure lessens and the build up stops. This then causes the layer to contract and the cycle restarts.

A demonstration of the effect this has on the temperature of the star is shown in Figure 1.7, where the surface temperature varies as the object does - matching the light curve shape of the Cepheid strongly.

There are some circumstances where this can change - either through natural evolution of the star altering the balance between the instability and damping forces, cutting the oscillations or driving them further.

The layers that are pulsating vary between stars but are usually partially ionized hydrogen or helium. The behavior of the layers depends on composition, density and temperature, hence why certain conditions must be met in order for their to be oscillation rather then the system balancing out [Gastine & Dintrans 2010].
What makes the group so valuable as a scientific tool and subject of study is how well most of the above categories (all bar ROap and Gamma Doradus) hold to relations between parameters. One of the most important discoveries in modern astronomy came when Henrietta Leavitt discovered a relationship between the period and luminosity of Cepheids \cite{leavitt1912}, known as the Leavitt Law. It was discovered when, looking at Cepheid variables in the SMC, where all the Cepheids were assumed to be roughly equally distant, that there was a simple relation between the brightness of the objects and their periods. Her initial plot is shown in Figure 1.9. This relation allowed astronomers to determine the cosmic distance to distant objects, aided by the intrinsic brightness of Classical Cepheids and their regular variations. Its use as a distance indicator was vital in the work of \cite{hubble1953} on the distance ladder and eventual creation of the Hubble constant.

A proper understanding of variable stars is especially relevant here - some initial estimates of the Hubble constant put in as high as 500 $kms^{-1}Mpc^{-1}$ \cite{hubble1929}, due to difficulty distinguishing between type II and classical Cepheids. The two have quite different relations between their periods and luminosities, as can be seen in figure 1.8 leading to very suspect results.

As the number of variable stars detected increased, similar trends were observed between other Cepheid-like variables and with the advent of large scale surveys of the sky, these relations have been more precisely measured over time, both for

---

**Figure 1.7** Cepheid CD Cyg, showing its variation in temperature (left) and log gravity (right) over one cycle. The red points denote measured values whereas the blue line denotes calculated values. Authors note that the large log g variations are probably due to bad decomposition of the line profiles used to measure the gravity. Figure 1 from Rastorguev et al. (2019)
Figure 1.8  Period-Wesenheit (a reddening-free quantity used as a substitute for magnitude) relations for Classical Cepheids, type II cepheids and RR Lyrae. Figure 7 from Soszyński et al. (2011b).

Cepheids (Riess et al., 2016; Macri et al., 2015; Soszyński et al., 2015; Matsunaga, 2013; Groenewegen & Jurkovic, 2017), RR Lyrae (Catelan et al., 2004; Bono et al., 2003; Marconi et al., 2015; Neeley et al., 2017), or other smaller categories like Delta Scuti or SX Phoenicis stars (Poleski et al., 2010; McNamara, 2011; Navarrete et al., 2017). The power of the Leavitt Law and similar relations as well as aiding us in understanding the later stages of evolution of these low to intermediate mass stars, which means finding these objects has a high priority. These types of variables are of such scientific value that some variability surveys use pipelines specifically to validate and characterize Cepheids and RR Lyrae specifically (Clementini et al., 2018).

As a general rule for these stars, the PL relation (or equivalent) displays some scatter in the optical bands, but in the near-IR to mid-IR, the scatter around the relation is decreased, and in some cases, like Type I Cepheids, any dependence on metallicity is almost completely removed (Bono et al., 2010). An example of this tighter spread in PL relation is shown in Figures 1.11 and 1.12 where the NIR observations have a much tighter spread, allowing the relation to be more accurately drawn. This scatter is partially caused by the finite width of the instability strip, but the IR surface brightness is less sensitive to the instability strip, causing less scatter about the relation (Wang et al., 2018; Madore &
Figure 1.9 Original plot from Leavitt & Pickering (1912). Apparent Magnitude vs log(period) for Cepheid variables in the Small Magellanic Cloud. The two distributions correspond to the maxima and minima of the objects.
Figure 1.10  *Slope (and associated errors) of period-luminosity relations of LMC cepheids.* The green points are theoretical model and yellow are the observed data. As can be seen, at longer wavelengths the relation has less scatter. Figure 1 from Subramanian et al. (2017).

This does not mean the amplitude of variation in the IR is higher (Subramanian et al., 2017), but there are other benefits to observing in near-IR that compensate for this, as can be seen in Figure 1.10. This topic will be covered further in Section 2.

This a particularly relevant topic here because out of all the categories of variables listed here, observations of pulsating variables are affected to some degree by wavelength observed at. Pulsating variables are the variables chosen for deeper study, along with eclipsing variables, due to their number and scientific applications. Several previous studies have looked into the wavelength dependency of variable stars but the focus has always been on the most thoroughly studied variable stars - RR Lyrae (Madore et al., 2013) and classical Cepheids (Madore & Freedman, 2012). However, little information is available into the wavelength dependence of any other objects, such as Long Period Variables.

The Long Periodic Variables (LPVs) are a closely related group of stars that contains Miras, Semi-Regular (SR) variables, OGLE Small Amplitude Red Giants (OSARGs - first named by Wray et al. (2004)) and Slow-Irregular (L) variables. They are a group of giants, bright giants and supergiants and consist of very red
Figure 1.11  Period-Luminosity relation for a sample of Gaia DR2 cepheids in the V Band. Part of Figure 4 from Lazovik et al. (2019).

Figure 1.12  Period-Luminosity relation for a sample of Cepheids using the four WISE bands. Black dots are fundamental mode classical Cepheids and red crossed are first-overtone classical Cepheids. Figure 2 from Wang et al. (2018).
stars that occupy a region of the HR-diagram quite separate from the Cepheid-like variables, to the right of the instability strip in Figure 1.5. Many of the Miras are some of the longest known variable stars ever found, with many discovered before modern astronomy due to their changes being visible to the naked eye.

These stars present a different challenge to detect than Cepheids. The shortest periods detected are between 10 to 100 days, with typical amplitudes of 0.005-0.13 mag in I band (Wray et al., 2004) for OSARGs. For the rest of the group, periods can reach up to thousands of days with the brightness increasing by as much as 1000 times. This requires sufficient observing time of target objects in order to approach the period of these objects and confirm their identity as a periodic variable and not an outburst.

The various groups were all observed to fit a range of Period-Luminosity relations (Wood et al., 1999). The LPV’s form a interconnected series with each other as a track of stellar evolution, which Soszyński et al. (2013b) justify using the connectedness of the group labeled C in Figure 1.13. This group contains both Miras and SR variables, with SR variables usually pulsating with two modes of oscillation, and thus typically occupying both the C and C’ relations. Soszyński et al. (2013b) argue that as the LPVs age, they smoothly transition to pulsating as Miras.

During a star’s first evolution up the Red Giant Branch, the star pulsates as an OSARG leading to the shorter scale variability (Soszyński et al., 2004). As the star continues to expand and brighten the amplitude of its variations increases as the period of oscillations decreases with the stars increased size, with tenths of a magnitude variations increasing to whole magnitude variations as the star moves along the Asymptotic Giant Branch, finally reaching the peak and pulsating as a Mira, with their impressive fluctuations which can reach up to 8 magnitudes in the (Soszyński et al., 2009; Sterken & Jaschek, 2005). This pulsation occurs due to the inherent instability of this period of a stars evolution. As K and M type giants, the typical (starting) mass of an LPV is 0.5–8 $M_\odot$ (Höfner & Olofsson, 2018). The large size and low mass result in surface velocity well below that of a sun-like star, resulting in heavy mass loss $10^{-7}$ to $10^{-6} M_\odot$/yr (Decin, 2020) at relatively slow speed (Arndt et al., 1997).

LPVs are very common variables (with OGLE observing over 200,000), due to most RGB and AGB stars showing intrinsic variability (Soszyński et al., 2013a) and form a key part of this project. Being far more common than Cepheid
Figure 1.13  Period-Luminosity relation for variable red giants in the LMC. RGB OSARGs are light blue, AGB OSARGs are dark blue, Miras and SRV are pink if O-rich and red if C-rich. The green points are stars with long secondary periods. Figure 1 from Soszyński et al. (2007).
and RR Lyrae, these variables represent an underutilized technique for distance measurement. There have been relatively few studies using LPVs for distance determination and comparing their performance with the more well-established pulsating variables, but the few that do exist have reported positively. Rau et al. (2019) examine LPVs and Cepheids in the LMC and produce a period-luminosity relation with comparable scatter to that of Cepheids for a restricted subsample of OSARGs. Selecting only their best fitting 20% of objects still provides them with more variables than the number of Cepheids with better reliability against extinction due to their colours. For these reasons, LPVs are the primary targets for this project.

Blue-White variables consist of upper main sequence pulsating variables, which contains Beta Cephei (Kahraman Aliçavuş & Handler, 2020) and Slowly Pulsating B (SPB) stars (Catelan & Smith, 2015e), and blue supergiant variables, which includes PV Telescopii, Alpha Cygni variables, Blue Large-amplitude Pulsators (BLAPs) and S Doradus (sometimes referred to as Luminous Blue Variables or LBV).

They are all massive, hot stars (Catelan & Smith, 2015f), being spectral type B to O (depending on the type of variable). There are exceptions to this, such as S Doradus stars which cool and reheat dramatically over the course of their cycles (Clark et al., 2005). These stars are all of great scientific interest but will not be the key focus of this project which will concentrate on cooler variable stars. These objects have high temperatures and rare due to their high masses and short lifespans although the are very luminous and can be seen over great distances if extinction is low.

The next category is that of Hot Subdwarf Pulsators. These stars exist in a region between the Zero Age Main Sequence (ZAMS) and the white dwarf cooling region of the HR diagram (see Figure 1.3). Due to their temperature and lower luminosity than MS stars of the same temperature, they have been classified (Moehler et al., 1990; Drilling et al., 2013) as Hot Subdwarf stars.

In general, these stars are low mass (on average 0.5 solar masses) Horizontal Branch stars, with effective temperatures between 22000 K and 40000 K (Heber, 2009). They are the aftermath of the core helium flash event in low mass stars (Heber et al., 1984; Saffer et al., 1994), with their temperature making them O to B spectral type stars. Four subgroups of hot subdwarf pulsators exist: sdBV_p, sdBV_g (or ”Betsy” variables), sdOV and He-sdBV variables, split by composition.
and restoring force - pressure or gravity (Catelan & Smith, 2015d).

The final group of pulsating variables to consider are pulsating degenerate stars. This contains stars along or near the cooling curve for white dwarfs, which are typically labeled as GW Vir, DAV (also known as ZZ Ceti throughout a lot of the literature), DBV, DQV and GW Lib variables (Vauclair, 2013).

1.2.2 Extrinsic: Eclipsing

Eclipsing Binaries are one of the most common types of variables objects due to the high fraction of stars that form as binary or higher-order star systems, between 40% for M dwarfs to 70% or more for A and earlier stars. (Kovaleva et al., 2019; Kouwenhoven et al., 2009; Avvakumova et al., 2013). The variation occurs in binary (or more) star systems where the orbital plane is inclined such that the component stars periodically eclipse one another, causing a reduction in amount of light an observer detects. Eclipsing Binaries can be stars of a variety of types which makes them useful scientific tools, from being allowing the measuring of masses and radii (Torres et al., 2010; Zhang et al., 2022) to being the test labs for gravitational waves, although this also adds slight complications from the much wider range of parameter space that they occur in, but using the fact that the stars will be of the same age, metallicity and distance can help to constrain the issue.

These types of variables are useful from a scientific perspective, allowing study of the sizes and distance between the two stars. For "detached" binaries, accurate estimates of the radius and masses of the stars can be made (Andersen, 1991). They can also serve as distance indicators, which makes them a very important group to study. The sizes of stars in contact binaries can be measured from their period, and with color information, an estimate of absolute magnitude can also be made (Rucinski, 1997, 1994). This allows the star to be used as a distance estimator. This method is well tested, and has been used to calculate distance to the SMC with good accuracy (Graczyk et al., 2014).

Eclipsing Binaries exist in a variety of types, caused by factors including component star types, component separation, roche lobe filling and the shape of the star (due to gravitational distortion). These types include:

EA-type eclipsing binaries (historically referred to as Algol-type variables) show
little change outside eclipsing events. If the stars have a large brightness disparity then the dimming phases of the brighter star obscuring the dimmer may be hard to observe. The stars are not in contact with each other ("detached" or "semi-detached") and thus are spherical in shape. EB (Beta Lyrae) Systems are close enough to distort each other and to have mass flow between the stars. They have light curves that lack a constant period between the dimming phases and thus it is hard to distinguish a defined start or finish for the eclipsing events. This type is defined by a different amplitude change for the different stars eclipsing each other \cite{Samus2017, Malkov2013}. EW (W Ursae Majoris) systems are close enough be in partial contact or full contact with each other and have similarly constantly varying, roughly sinusoidal light curves as EB binaries. This type has each minima from the different star obscuring the other being roughly equal depth \cite{Samus2017}. For an example of what visually distinguishes the light curves of these types, see Figure 1.14.

The next set of categories group stars by physical characteristics, such as stellar type. The General Catalogue of Variable Stars (GCVS) describes multiple categories for giant and white dwarf stars as well as stars with strong sun spot activity (RS Canum Venaticorum) \cite{Samus2017}. The last set of categories encapsulates the degree of filling of the stars roche lobe or lobes, from "detached" systems where both roche lobes are not filled to "contact" systems where the stars are in contact with each other \cite{Samus2017}. Many of the labels used for eclipsing binaries and somewhat even variable stars in general are subjective, based on observed properties, which can change from instrument to instrument and over time. The common labels used here for eclipsing binaries are often overlapping and stars often have several, each assigned by different authors. It is important to understand properly what each classification represents.

With many kinds variable stars it is a possibility that a star might display multiple kinds of variability, for example a pulsating star showing characteristic signs of two separate kinds of pulsation each driven by a different mechanism simultaneously. Eclipsing Binaries are the most common way by which a star varies in multiple ways at once, such as a classical Cepheid companion, a Delta Scuti companion \cite{Gieren2015, Yang2018, Samus2017}, see Figure 1.15.

These stars are extremely common, but the additional compound variation can lead to difficulties in determining which behavior is due to which method, especially with similar periods for each mode or spotty light curve coverage. Additionally, these objects do not vary as dramatically as some other variable
Figure 1.14  Sample light curves for different types of eclipsing binaries from VISTA. The spotted type refers to eclipsing variables such as RS Canum Venaticorum, where the presence of sun spots can have an additional noticeable effect. Figure 13 from Angeloni et al. (2014).

Figure 1.15  I band model of overlapping variability during eclipse of a Cepheid. Bottom plot is light curve and model, upper plot is residuals of the model. Object measured by OGLE (Pilecki et al., 2013).
types, with low amplitudes that can be overlooked. Finally, any process which is atmospherically based or heavily involves a physical change to the outer layers of the star, will by it’s very nature, be impacted by the composition of the stellar atmosphere and how it’s absorption features interact with different wavelengths of light. The spectra of certain stars may result in the variations seem in certain wavebands being dampened by changes in the outer regions of the star or matter surrounding it. Eclipsing stars are not immune to this effect - with the presence of expelled material via stellar wind or accreted from another source, they can be affected in the same manner as pulsating variables.

1.2.3 Extrinsic: Rotating

The apparent brightness of rotating variable stars changes as the star rotates, either because of surface features of the star that would result in different surface brightness (such as concentrations of sunspots) rotating in and out of vision (Percy, 1978), and/or because of rotation causing the star to distort into an ellipsoidal shape, resulting in differing brightness as it rotates and different surface areas of the star are visible. Figure 1.16 from (Eyer & Mowlavi, 2008) demonstrates the position of (some) rotating variables on the HR diagram.

Alpha$^2$ Canum Venaticorum (ACV) variables and Rapidly oscillating Alpha$^2$ Canum Venaticorum (ACVO) variables have changes driven by peculiarities in their magnetic fields, and are main sequence stars with spectral types B8p-A7p (Catelan & Smith, 2015b). They are chemically peculiar stars which show unusual metal abundances in their upper layers (Samus' et al., 2017). SX Arietis variables are similar, being peculiar main sequence B-type stars, typically hotter than ACV variables (Samus' et al., 2017).

Other rotating variables are binary star pairs. Rotating ellipsoidal variables are binary star systems that do not eclipse each other like eclipsing variables but show light variations from changes in the emitting area visible as the pair rotate and their distorted areas rotate in and out of vision. FK Comae Berenices variables are spectral type F and G giant stars and particularly rapid rotators, to the point of appearing ellipsoidal. This extreme rotational velocity is suspected to be caused by recent contact binary mergers (Ayres et al., 2016).

BY Draconis and RS Canum Venaticorum are both of interest because they show multiple types of variability quite commonly. BY Draconis variables are K to M
Figure 1.16 Color coded HR diagram using variables from the Hipparcos catalogue. Figure 4 from Eyer & Mowlavi (2008)
dwarf stars on the main sequence. It is believed that starspots are the reason for their variability \cite{Holl2018, Lanzafame2018}. They typically have amplitudes of 0.5 mag and periods of a few tens of days, although they are irregular and can change periods over time. They can also display flares, characteristic of eruptive variables of the type UV Ceti, described in the next section. RS Canum Venaticorum (RS) variables are binary stars that can show eclipsing, rotational and eruptive variability. They are believed to have starspots on their surface like other rotating variables, can show eclipses as the pair obscure each other, and eruptions if one of the members is a flare star \cite{Lanzafame2018} or an accreting white dwarf. The stars are typically F to K subgiant and giant stars (see the Gaia Collaboration’s paper \cite{Lanzafame2018}, Figure 2 containing a sample light curve and how they selected these objects).

Rotating variables are fascinating glimpses of stellar structure, so any found may be interesting probes into regions hard to measure normally. As a downside, these objects are relatively uncommon (compared to OSARGs or eclipsing binaries), often low amplitude \cite{Lanzafame2018b} and even the more common occurrences share similar light curve shape and regular periodicity with other types of variables.

### 1.2.4 Intrinsic: Erupting

Erupting variables is a broad category that describes several types of stars that vary irregularly or semi-regularly primarily due to outbursts of material from the star. They can be divided into two rough groups.

The first group is mainly young stars or proto-stars, such as T Tauri and Herbig Ae/Be stars (HAeBe) - sometimes referred to as Orion or Nebula variables, the latter name given by their common environment.

T Tauri stars are stars below about 2 solar masses moving along the Hiyashi track (see \ref{1.17}) towards the main sequence. They are F, G, K or M spectral types and luminous due to their large sizes. Herbig stars are between 2 to 8 solar masses and spectral types between F2 and O9 \cite{Donehew2011}. T Tauri are primarily found in nebulae and both tend to have accretion disks \cite{Donehew2011} and vary dramatically through starspots, outbursts of stellar winds \cite{Waters1998} and obscuration from their disks and proto-planets and other objects forming in them \cite{Rodriguez2017}. Other Orion
variables include FU Orionis variables, which feature a large (∼6 mag (Jurdana-Šepić & Munari, 2016)) brightness increase and spectral type change. In nebula there is typically very high extinction, which can reach over 100 mags in the densest regions (Itoh & Oasa, 2019). The near-IR observations of VVV are a great benefit when dealing with these and other dusty objects.

The amplitude of such objects is large, between 0.5-3.5 \( \Delta K_s \) but varies heavily between objects (Medina et al., 2021). Eruptive stars can vary due to accretion events, typically resulting in large outbursts and rotation, spots or obscuration, leading to smaller changes. (Rice et al., 2015) group them into categories that they found best characterised the behavior observed and allowed for best separation: Protostars, Disked Sources and Non-Disked Sources. These three categories increase with age starting from Protostars and decrease in typical amplitude.

The technique of infrared excess is commonly used for probing the presence of a disk or otherwise large quantity of circumstellar dust (Froebrich et al., 2021). Lower infrared colors suggest an unobscured stellar photosphere, whereas higher values indicate infrared excess.

These stars are also sometimes referred to as "IN" variables (Samus’ et al., 2017). This designation labels them as Irregular Nebula variable stars, but it is my opinion that these classifications are too broad and overlapping to be the preferred choice of labels for such objects. T Tauri are given the label INT. For other young stars found in nebulae, they are typically referred to as INA or INB variables if they are early spectral types (O-A) or late spectral type (F,G,K or M) respectively (Samus’ et al., 2017). If the light variation occurs on a timescale of a few days, they are labeled INSA or INSB. If these variables are detected not in a nebula, they are labeled as IA, IB, ISA or ISB.

The other main group of eruptive variables are large, usually evolved stars. Their size means that can lose mass easily to stellar wind and other outbursts or eruptions.

As mentioned in Section 1.2.1, S Doradus variables show small regular variations like other giant stars and larger eruptions (Drissen et al., 2001). They are referred to here as S Doradus variables and not LBVs following the nomenclature of van Genderen (van Genderen, 2001), who points out that all evolved stars are micro-variables, but only a few are S Doradus, making it a more sensible designation.

Other eruptive variables include Gamma Cassiopeiae (GCAS) variables which are
Figure 1.17  Evolutionary tracks of pre-main sequence stars, labelled by mass. The Hiyashi track is visible as the vertical paths undergone by lower mass stars. Image taken from Figure 2 of Palla (2012).
B giants and sub-giants that rapidly rotate and eject matter from their equatorial regions, R Coronae Borealis variables are supergiant F and G stars that show regular pulsations and large drops in the visual bands (Rosenbush 2008) due to formation of carbon dust in the stars atmosphere (Rosenbush 2008). This makes the star a very interesting prospect to observe in IR where the magnitude change is smaller, although this type of star is rare (Rosenbush 2008). BE variables behave similarly to GCAS variables, with rapid rotation forming a disk around themselves. For a review of these interesting emission line stars, see (Rivinius et al. 2013). Wolf-Rayet stars are extremely evolved massive (>10 solar masses) originally O-type stars (Crowther 2007), with extremely strong stellar winds that can form a large bubble around the star (Dwarkadas et al. 2018). These stars are incredibly interesting testbeds for stellar evolution, especially at the end of a star’s life (Schootemeijer & Langer 2018), and the chemical enrichment of galaxies (for a review, see Crowther (Crowther 2007)).

The last category is Flare stars, otherwise known as UV Ceti variables. They are M and K dwarf stars that have strong outbursts for brief periods due to the dwarfs active magnetic field. They can also be observed in stars of most other types across the HR diagram, including solar-like stars (Balona et al. 2016) and are thought to be particularly strong outbursts in the same manner as the Sun’s solar flares.

Eruptive stars are irregular, but many of the younger eruptive variables - often grouped together into the category of YSOs (Young Stellar Objects) are heavily dust shrouded. VVV is able to observe these objects in greater clarity then many visual surveys are capable of, and as such, I expect to detect them relatively frequently.

### 1.2.5 Intrinsic: Cataclysmic

This category of variable objects all have a common source. These objects originate from matter accreting onto a compact object - usually a white dwarf. Given the underlying process, these objects are typically binary systems containing a white dwarf and a partner star or other source of accretable material. The variability results from the (typically giant partner star) companion transferring mass to it’s companion, triggering an outburst (Warner 1995). This can occur in several forms.
Novae usually have a partner star of spectral type from K to M (Samus' et al., 2017). Over the course of a few hundred days the system brightens by between 7 to 19 mag (Samus' et al., 2017). There are several different subcategories including fast novae (NA) which are steeper and last less than 100 days, slow novae (NB) and very slow novae (NC) which rise and fade over the course of many years (Hoffmeister et al., 1985). Recurrent Novae (NR) are when multiple novae have been observed from the same binary pair, typically many years after the first (Hoffmeister et al., 1985). These classifications are based on observational features, which may or may not be strongly connected to the underlying physical principles. There is likely much overlap with the following categories of cataclysmic variable, with some positing that such stars will vary between novae and dwarf novae throughout their life (Shara et al., 2007).

Dwarf Novae (also known as U Geminorum variables) are binary systems with a subgiant or dwarf companion. Compared to Novae, Dwarf Novae typically show a smaller variation over a smaller period of time, and typically future eruptions happen sooner than for Novae. The group is divided into subgroups, again by physical appearance or light curve shape: UGSS (named after the prototype star SS Cygni) variables increase by 2-6 mag in the V band over a few days and return to their original state shortly afterwards (Samus' et al., 2017). UGZ (named after Z Camelopardalis) variables are similar to UGSS variables, but have "standstill" phases where they remain at approximately constant brightness roughly between maximum and minimum (Osaki, 1996), believed to be due to a large and steady flow of accretion material - too large to produce a dwarf nova outburst but not enough to trigger a full novae (Oppenheimer et al., 1998). This does however provide fuel for a period of consistent bright phases. The third category are UGSU (SU Ursae Majoris) variables. These are characterized by outbursts of over 6 magnitudes, most with similar outbursts to normal Dwarf Novae, but some without (Osaki, 1996).

There are a few others which are broadly generalized into Symbiotic Novae or Z Andromedae variables. They arise from a hot evolved object and a large cool partner, with the mass transfer occurring through stellar wind from the cool giant. The fairly steady outflow of wind leads to long periods of maximum as the built up wind is consumed and a prolonged period of quiescence as it builds back up (Merc et al., 2019; Sokoloski et al., 2006). SW Sextantis variables experience very heavy mass transfer, are similar in light curve shape to eclipsing nova-like stars and have characteristic single-peaked emission lines (Schmidtobreick, 2017). VY
Sculptoris variables are posited to have the white dwarfs magnetic field strongly influence the incoming mass flow, leading to low or completely stopped accretion (King & Cannizzo 1998).

Supernovae are an extremely bright type of Cataclysmic event, with several possible causes and a resulting explosion that completely disrupts the star (Woosley & Weaver 1986). They are extraordinarily bright, rare, easily distinguished by their rapid change in brightness and timescale of decline and well known enough that they will not be exhaustively covered here.

Because a lot of the luminosity that makes Cataclysmic variables detectable comes from the White Dwarfs accretion disk and material being transferred from it, they are traditionally observed in the optical and ultra-violet, which is where the spectrum of accretion generated luminosity typically peaks at (Hoard et al. 2002). Because of this, there have been relatively few surveys for Cataclysmic variables at longer wavelengths. Because these objects can vary in brightness over short periods of time, it is mainly with the advent of large scale sky surveys such as The 2 Micron All-Sky Survey (2MASS) and Wide-field Infrared Survey Explorer (WISE) that studies on a large scale can be done (Hoard et al. 2002; Evans et al. 2014). Previous IR study of Cataclysmic variables were largely done in response to an individual report of a Nova (Evans et al. 2014).

Whilst the study of such objects of the scale of VVV is an interesting idea to detect and learn more about this type of variables in dusty regions where previously they were unseen. The cataclysmic variables in general are not amenable to the methods that we will use to classify variables. The long periods of quiescence and (usually) short bursts of activity is not an ideal match with VVV’s sporadic scanning. Any objects that happen to be found will be interesting, but cannot be relied upon, so further discussion of them will be limited.

1.2.6 Intrinsic: X-Ray Bright

These variables are defined by strong X-ray sources, typically resulting from the innder part of the accretion disk of a compact object such as a neutron star from a binary partner star (Taranova & Shenavrin 2017). X-ray variability surveys do exist, both extra-galactic and intra-galactic, searching for Active Galactic Nuclei and x-ray binaries inside our own galaxy. Although they are mainly detectable from X-ray emission, there is a non-negligible (∼ 1 mag in the JHK near-IR
bands (Taranova & Shenavrin, 2017)) amplitude component emitted in the near-IR bands as well from accretion disks or shell emission in the case of (Taranova & Shenavrin, 2017). Despite them being quite uncommon (Debosscher et al., 2007), they do share a lot of overlap with the Catalysmic group of objects. The same principles apply here: any occurrences detected will be noted, but they will not be the primary targets.
Chapter 2

Variable stars and their properties

With the wide variety of variable stars and increased coverage and depth of modern surveys, there has been an increasing trend in the number of variable objects observed. In the space of ~20 years, the EROS (Experience pour la Recherche d’Objets Sombres) survey went from a few hundred variable stars (Grison et al., 1995; Afonso, 1999) to ~60,000 (Kim et al., 2014). We have new studies like the Catalina Surveys Periodic Variable Star Catalog (Drake et al., 2014) which contains ~61,000 variable stars, the All-Sky Automated Survey for Supernovae (ASAS-SN) (Pawlak et al., 2019) finding ~220,000 variable objects and the Gaia DR2 variability release (Holl et al., 2018b) contains ~550,000 variable stars. As more and more are found, this leads to greater understanding of the different types of variable star. As discussed in the background in Chapter 1, there are many reasons behind variability - with internal or external factors responsible. For internal processes, whilst there are strong theories about what physical processes drive variable behaviour, there remain issues with how stars evolve to this state or what triggers the phenomenon. With types of variables whose variability is caused by external factors, the question of how the star evolves to that state is less relevant then how the system forms.

As touched upon in the background there are a very wide variety of variable stars, some with properties unique to them and some that share features in common with other types of variables. As mentioned previously about the mix-ups that can occur when trying to use the properties for one type of objects on another of different type - such as Period-Luminosity relations, this is a very important issue to be able to both understand and adjust for.
Due to similar processes, many light curves resemble each other between groups. In addition, we only observe the object sparsely - capturing only a representation of it’s true behavior. When hunting for periods or models the light curves of such objects, many systems default to sinusoidal or other oscillatory patterns. It can prove difficult to distinguish between similar oscillatory light curves if all other properties are similar, especially with the naked eye, and even computational methods can struggle when data is sparse. The lightcurve shape, period, amplitude and the objects colours can all overlap and cause confusion.

Additionally, whilst some variable objects may be straightforward as to how they arise (such as eclipsing binaries), others factors at play in observing such objects - such as uncertainties in their environments due to dust, uncertainties in measurements of properties such as distance or mass add additional complications.

More objects being found means we are more likely to uncover new or unusual circumstances - such as the recent strange observed activity of Betelgeuse. The red supergiant dimmed by 35% and then recovered to its previous brightness over several months at the start of 2020 (Guinan et al., 2020). Instruments were scrambled to observe and the current best conclusions reached are that a newly formed dust clump obscured the southern hemisphere, caused by cool ejected matter from the photosphere of the star (Montargès et al., 2021), a local photospheric cooling (Taniguchi et al., 2022), or a combination of the two (Cannon et al., 2023). By observing more and more variable objects overall and improving our candidate pool, we can use the new objects to improve our understanding.

### 2.1 Targets

While all variable stars are interesting, it would not be plausible to direct equal and comprehensive attention to all of them.

Some variable stars change on timescales of hours to minutes and without observations at cadence close enough to that timescale, we cannot see such structure as it occurs. Other variable objects vary unpredictably and erupt in flares (Benz & Güdel, 2010) or novae (Kalomeni et al., 2016). Finding these objects is better carried out by surveys with rapid follow up systems such as astronomical alert brokers (Förster et al., 2021), whereas VVV requires several weeks to perform its analysis on new data. Some objects that burst into visibility
before fading again can still be detected if the variability is repeating, even if not periodic, such as dwarf novae (Szegedi et al., 2022). For non-repeating objects, any occurrences of such objects is mostly chance and cannot be relied upon and follow up observations will only exist long after the original event. These objects deemed as unfeasible main targets includes the eruptive, cataclysmic and many of the x-ray bright objects, which fall into the second of the categories above. Pulsating white dwarf stars fall into the former category and are also unfeasible.

This leaves several different categories of variables remaining: Eclipsing, Rotating and most if the Pulsating variables. For my project, I have decided to focus on part of the pulsating variables - Long Period variables. These low-mass and evolved stars are red and can be dusty due to mass loss - a problem for optical observations but less of an issue for VVV’s near-IR observations. As pulsating variables, LPVs have period-luminosity relations, but are largely measured in the optical only (Soszyński, 2022; Iwanek et al., 2021b). With VVV data, I can expand our coverage of these relations.

The remaining variables cannot be ignored however - Cepheid-like variables are some of the most well studied variable stars and fall into the same broad class of variable star as Long Period variables - Pulsating, sharing similarities in variability process and variation structure. Part of the other reason to work with cepheids as well is to remove contaminants. When working with objects that can exist in many different forms and states (some even with overlap), there can be substantial difficulty distinguishing between types of objects (Christy et al., 2022), especially en mass, where looking at each individual light curve or pursuing followup observations isn’t possible.

Eclipsing Binaries are by far the most common type of variable and cover a broad range of types and morphology’s (Petit, 1987; Hajdu, 2021; Catelan & Smith, 2015b) and so any study on variables that did not account for a fraction of any list of candidates to involve binary stars, runs the risk of falsely classifying binary stars as the variables of interest.

Rotating variables do not share the common morphology with LPVs and are not as populous as eclipsing variables, although they are not rare either - large variability surveys such as ZTF have found roughly as many of these variables as they did Pulsating ones (Chen et al., 2020). RS CVn variables are also a type of eclipsing binary and can be difficult to separate from non RS CVn binaries without access to the objects luminosity. BY Dra rotating variables are dwarf
stars with similar periods of RS CVn variables. Both are low amplitude and short period (Chen et al., 2020), which makes them not ideal targets, but they are common enough that the VVV will have detected a sample of them.

When dealing with objects with similarities, be that through evolution or morphology, there can be common factors between the types. To better understand the problem, we must understand what the target objects are and what are the factors that could complicate identifying such an object.

### 2.1.1 Target Details

In addition to the different types of factors separating the Long-Period variables from the other types, we must also deal with any confusion between the different Long-Period variables themselves. As per the name, part of their main identifying factor is their period of variation, which, when combined with amplitude is used to drawn the boundaries between the different types.

A Long Period Variable (LPV) star in this case refers to Miras, Semi-Regular Variables, OGLE Small Amplitude Red Giants and all unclassified LPV stars.

The boundaries to differentiate, especially between Mira and SRV are somewhat subjective. It is usually agreed upon that Mira variables are defined by having a large (>2.5) optical amplitudes and that SRV stars have smaller amplitudes and oscillate in more excited pulsation modes (Iwanek et al., 2022b). This introduces another complication which will become relevant - the amplitude and period can vary depending on what wavelength one is observing at.

Semi-regular variables are defined by optical amplitudes of >1 mag and periods of 50-100 days. As per the name, they are a somewhat loosely defined group with frequent occurrences of multiple periodic sequences occurring in light curve observed from the star.

Lastly are OSARGs - OGLE Small Amplitude Red Giants and the newest category I will be looking at. Their newness and lack of understanding and other research done on such objects has made them one of the types of variable star I am focusing the most on. They are defined by their small amplitude - <0.1 mag in the V band, and periods from 10-100 days.

All of these stars lie in similar regions of their evolutionary history. They are
Figure 2.1  Example colours of Mira with known spectral types (top) and all Mira (bottom) present in the Kilodegree Extremely Little Telescope (KELT) survey. In the top plot O-rich stars are shown as blue circles, C-rich as red circles, and S-type in orange crosses. The plots compare the $J - K_s$ 2MASS colours with the Wesenheit index calculated with the 2MASS $J$ and $K_s$ magnitudes and the $G_{RP}$ and $G_{BP}$ the Gaia magnitudes. Figures 11 and 12 from Arnold et al. (2020).
Figure 2.2 Comparing the amplitudes and light curves of LPVs compared to eclipsing binaries. The image shows phased lightcurves for an eclipsing binary (left) and an LPV (right). The two horizontal lines are magnitude limits used by the authors to aid in picking out eclipsing binaries. Figure A.1 from Groenewegen (2005)

relatively low mass objects undergoing their red-giant, asymptotic giant branch or post asymptotic giant branch stars, leading to them being quite red and relatively cool. Example colours of Mira are shown in Figure 2.1 - will all of them having positive (J - Ks) colours and a large majority - the O-rich Miras - having near-IR based Wesenheit index ($W_{Ks}$) greater than the optical based index $W_{RP}$. The Wesenheit index is a reddening free index calculated as $W_{Ks} = Ks - 0.67(J - Ks)$ and $W_{RP} = G_{RP} - 1.3(G_{BP} - G_{RP})$ where J and Ks describe the 2MASS magnitudes and $G_{RP}$ and $G_{BP}$ the Gaia magnitudes.

There are even issues distinguishing between using their colours. The amplitude varies depending on what wavelength one observes at - for us, with our near-IR observations, the large Mira amplitudes observed in visual (4-8 mag), reduced down to 1-3 mag in the Ks band.

The period is another strong distinguisher, but suffers from being computationally expensive to calculate on large scales and can be heavily dependant on the quality of the data itself and the number of observations plays a key role in whether or not one can even find any periodic information at all.

As mentioned before, the other types of variable star can also complicate matters. Very few have amplitudes as high as Mira variables in the optical, but many including eclipsing binaries have similar amplitudes in the near-IR (Gramajo et al. 2020; Whitelock 1998) since many types of variable do not have as strong wavelength dependence on amplitude as Miras. As can be seen in Figure 2.2, eclipsing binaries show strong amplitude variation but have very different light curve shapes to LPVs. The two objects have completely different sources of variation, the light curve is one approach for tackling the problem of separating
out the two. Eclipsing binaries can show a wide range of periods - from incredibly short oscillations caused by very close contact objects to binary systems with a great distance between the objects.

For OSARGs, their small amplitudes and short periods 1-50 days (Rau et al., 2019) place them in overlap with many other pulsating stars such as cepheids. Whilst some of these stars have quite distinctive light curve shapes with distinctive peaks, many such features fall away at longer wavelengths, making them less reliable tracers. Examples of OSARG light curves are show in Figure 2.3 demonstrating how even between objects of the same variable type, there can still be strong variation between the light curves of individual objects. The figure also demonstrates the overlap between light curve shape that complicates distinguishing a single object - the fourth OSARG has a strongly similar light curve to a SRV.

The OSARGs redder colours would be sufficient to separate them from eclipsing binaries which are usually hotter and bluer stars, but because VVV is surveying such a obscured region, there is a substantial problem with reddening. Without access to high quality extinction maps it can be difficult to tell between an OSARG with little extinction and a eclipsing binary with heavier extinction.

Mira variables are low to intermediate mass stars on the asymptotic giant branch with periods from 100 days to over 1500 days and amplitudes of variability of greater than 2.5 mag in V (Whitelock, 1998), greater than 0.8 mag in I (Whitelock, 2012) and greater than 0.4 mag in K (Whitelock, 2013). Cool (surface temp \(\sim 3000\) K) and giant (Trabucchi et al., 2021a), Miras can be divided into two types: C-rich and O-rich with slightly different properties like many other AGB stars (Cioni et al., 2001).

The two types have different period-luminosity relations with O-rich Miras having tighter scatter by a few percent (Huang et al., 2018; Ou & Ngeow, 2022). The relationship is not linear and models comprised of two slopes or quadratic based models have both been used (Ou & Ngeow, 2022). It is not clear which of the two is superior at this point in time.

Whitelock et al. (2006) note that it is more difficult to differentiate between C-rich (carbon-rich) Miras and other long period variables than for O-rich (oxygen-rich) Miras. For the latter, they noticed that the JHKL colours differed between Miras and non-Miras (Whitelock et al., 1995) whereas the colours were more uniform for C-Miras. The low number of evolved C giants in their sample leaves the degree
Figure 2.3  Examples of OSARG I-band phased light curves (top 4 images) compared to phased light curves for LPVs and ellipsoidal variables with red giant components (below). Top 5 images sourced from OGLE team (Soszyński et al., 2013a). Lower image sourced from Figure 1 in Soszyński (2022)
of applicability uncertain, however. 

Ou & Ngeow (2022) divide Miras into another category on top of C-rich or O-rich. Most Mira have light curves that oscillate in a basic sinusoidal fashion as seen in Figure 2.2. Exceptions to this behavior are dubbed non-regular Mira and pulsate with a sinusoidal component alongside a much longer term variation that may or may not be periodic. This longer term variation has been detected occurring from 1000 to 5000 days. It may occur on longer timescales, but on such long timescales there are observational difficulties. The non-regular Miras appear to be mostly C-rich Mira and it is speculated that circumstellar extinction due to a dust shell created by the AGB star may be an important factor (Whitelock et al., 2003).

Mira variables have sawtooth shaped radial velocity curves with typical amplitudes of 25 km s\(^{-1}\) whilst SRV have smaller amplitudes and smoother shaped curves (Lebzelter et al., 2005) but there have long been difficulties distinguishing the two types apart. Studies have found SRV stars with periods in the domain of Mira (> 300 days) and amplitudes beyond the threshold of 2.5 mags in optical usually used to draw the separation between the two types. As has already been discussed, Mira have been detected with irregular periodicity (Lebzelter et al., 2005). It is clear that the transition between the LPVs is not fully understood.

Semi-Regular Variables (SRVs) are a more heterogeneous class of objects that Mira. They are found on both the RGB and AGB and, despite the name of the group, often are truly periodic objects. Stars like these experience increasing mass loss as they evolve up the RGB and AGB, resulting in complex dynamics that amplify any uncertainties on stellar parameters (Chiavassa et al., 2019). Papers have demonstrated how in samples of SRVs the large and few convective cells that occur in the upper layers of the star can cause the visible photocentre of the star to vary and dominate the parallax errors (Chiavassa et al., 2018).

Several papers have suggested using pulsation periods (Trabucchi et al., 2021b) and pulsation modes (Arnold et al., 2020) as the more sound method for identifying SRVs apart from Mira. They identify most SRV as overtone pulsators, associated with more complex oscillations occurring inside the pulsating layers of the star (Arnold et al., 2020), whereas Mira are fundamental mode pulsating objects.

Additional complications can arise from stars with Solar-like oscillations. These objects do not share the same source as the three very noticeable categories
given above. (Lampens, 2021) describes a study of solar-like oscillations and other variability (such as Delta Scuti) in Eclipsing systems. With such systems providing masses and radii, there is concrete information about each object in addition to anything obtained from the observations of each companion.

Highlighting the study of oscillating red giant stars (in distant binary pairs, so that the oscillations are not dominated by tidal effects from the partner), astroseismological models and their predictions of stellar parameters of mass and radii can be compared to the predicted binary values. There is a strong link between the pulsation frequency at max amplitude and separations by existing astrosiesmology scaling relations. Binarity is a strong way to test these relations (Gaulme et al., 2016).

For RGB stars such as these, Astroseismology may be a way to understand the micro-variations present in these stars, which could potentially be expanded to the rest of the LPVs. As OGLE has demonstrated, there are a \( \geq 100,000 \) number of objects with the small-to-moderate variations observed in OSARGs. The variations of pulsating stars (Gaulme & Guzik, 2014) all can be linked to astrosiesmic processes inside the star and drawing upon that field of studying knowledge to understand how, if at all, the kappa-mechanism driven variations of the OSARG connect to the smaller variations of a "solar-like" oscillator.

The process to find such objects shares similarities with multi-periodic investigations - folding the light curve, subtracting the primary period based curve and performing a periodicity analysis, such as Fourier, on the residual.

### 2.1.2 Additional complications

There are additional complications beyond differences in the stars themselves. VVV is observing very crowded regions of the sky - along the plane of the Galactic Disk and much of the Bulge towards the Galactic Centre. This burdens us with two key problems: Extinction and Crowding.

**Extinction**

Because of thick dust towards the galactic centre, we can, if not handled carefully, lose our ability to distinguish the colors and magnitudes of our objects accurately. This dust exists in clumps, not as a smooth continuum, and thus accounting for
it requires careful modeling and careful application of those models. The effect of dust is not a consistent factor - it varies depending on the wavelength observed at and the amount of dust the light observed is passing through (Sanders et al., 2022). There are ways to probe the latter and as such, maps of the extinction observed from our perspective have been created.

These extinction maps come in 2D and 3D versions. The 2D maps describe the observed extinction as it varies with sky coordinate across the sky. 3D versions of said maps include the distance each measured extinction is taken at, leading to each sky coordinate having multiple distances and different amounts of extinction at each. The 3D versions are much harder to create because they require the objects used to measure the extinction to also have well-measurable distances as well, usually leading to restrictions in applicable area. For my uses, I considered the maps of Green et al. (2019), Lallement et al. (2019) and Chen et al. (2014). In order to use the 3D extinction maps distances will be needed - these are calculated using crossmatched Gaia parallaxes and discussed in Section 3.2.3.

The Bayestar19 maps of Green et al. (2019) use a combination of Pan-STARRS 1, 2MASS and Gaia DR2 data to map out the sky north of a declination of -30°, reaching several kiloparsecs out, varying depending on area. This map is the third iteration in a series, improving on the previous two by using the new Gaia parallaxes to dramatically improve distance measurements, especially to nearby stars. Newer maps such as this have begun to incorporate probabilistic approaches and techniques that infer the dust distribution and the stars along its sight lines. The distributions of the 799 million stars used in the creation of the map are shown in Figure 2.4. As can be seen from the figure, it is beyond 5 kpc where the number of stars, and thus the accuracy of the model, begins to fall off.

The Structuring by Inversion the Local Interstellar Medium (Stilism) (Lallement et al., 2019; Capitanio et al., 2017) uses 2MASS and Gaia DR2 data. The method estimates extinction to a ∼16 million stars with good Gaia parallaxes and then these distance-extinction pairs are inverted to create a 3D map of dust density. The technique uses the same principles used in an early paper (Lallement et al., 2014), but with a drastic increase in samples from the original ∼71,000.

The grid size of the map is not uniform. Within 1 kpc the increased number of sources provides a 25 pc resolution, which increases to 500 pc at the outermost regions of the map. The area covered by the map is a $6 \times 6 \times 0.8$ (kpc$^3$) volume with an effective max range of ∼3 kpc. This is a drawback compared to the
Figure 2.4  Distribution of stars in the Bayestar19 (Green et al., 2019) extinction maps, plotted in solar-centric Cartesian coordinates. The black dot denotes the galactic centre. The colour scale from blue to yellow is an arbitrarily assigned logarithmic scale of density. Figure 17 from Green et al. (2019).

Bayestar19 map; with the galactic centre being $\sim 8$ kpc (Gravity Collaboration et al., 2019) away, this map provides extinction coverage for less than half that distance. In Figures 2.5, 2.6 and 2.7 the coverage and calculated density of the Stilism maps can be seen.

The last map considered are the maps of Schultheis et al. (2014), Chen et al. (2013) and Chen et al. (2014), slightly older 3D map towards the galactic plane. While these maps may be older and do not have access to the newer advancements in distance estimation thanks to Gaia parallaxes, these make use of VVV, 2MASS and GLIMPSE-II data. The greatly improved coverage of the stars used for mapping at longer wavelengths, especially the wavelengths I will be using and the enhanced ability to see through the stronger extension regions make this map appealing. As can be seen in Figure 2.8, this map also reaches great depths towards the galactic centre and beyond which may be necessary for some of the bright AGB and RGB stars in my sample.

Many of these extinction maps are designed for visual band extinction, thus requiring some conversion to the wavelengths used by the VVV survey. The absolute value of the extinction is given by $A_V = R_V \cdot E(B-V)$, where is the visual
Figure 2.5  Dust density in the Lallement et al. (2019) extinction maps along the plane with the sun at the centre and the galactic centre to the right. The colour scale shows the extinction strengths. The dotted line is the limits are which the best quality resolution of 25 pc grid size can be achieved. Figure 9 from Lallement et al. (2019).
Figure 2.6  Dust density at different vertical planes off the galactic plane. Colour scale and black line are shared from Lallement et al. (2019).
Figure 2.7  Dust density at different vertical planes off the galactic plane. Colour scale and black line are shared from [2.5, Figure 11 from Lallement et al. (2019)].
band extinction, E(B - V) the colour excess and $R_v$ a medium dependent quality, derived from observational measures. Figure 2.9 shows the ratio of extinction at different wavelengths, which can be used to convert between them.

There has been some discussion in the literature as to the best extinction law to use and the best value of the parameter $R_v$. The typical value for a diffuse medium is 3.1 (Voshchinnikov, 2012), but there is much discussion about the correct value to use. The VVV survey has observed the bulge and disk, both of which differ quite substantially in structure and composition - leading to very different dust and extinction profiles.

For my purposes since supplementary information from surveys less capable of seeing through dust is being used, they will struggle at the densest regions where a higher value of $R_v$ would be a better fit. Since most of the candidate objects will have other survey counterparts using $R_v = 3.1$, to make our our corrections in line with theirs $R_v = 3.1$ is used. Future work may target specifically the thicker regions of the bulge, and use altered values of $R_v$ accordingly.

Taking the above and the choice of extinction map (Schultheis et al., 2014) into account may lead to higher uncertainties in the magnitude corrections the maps predict. I do not believe that this will prove too large of a problem however; objects which feature large corrections or corrections whose uncertainties reach
Figure 2.9  Compilation of IR extinction laws from literature. Red stars plot the extinction law towards the galactic centre derived by Fritz et al. (2011). Blue triangles are derived from stars towards the galactic centre by Rieke & Lebofsky (1985). Green squares were derived from red clump giants by Nishiyama et al. (2009). Cyan diamonds are galactic center extinction found by Gao et al. (2009). The remaining three symbols are extinction laws obtained via sight-lines not towards the galactic centre. The solid and dash-dot lines denote the model extinction curves - with $R_{v}=3.1$ and $R_{v}=5.5$ respectively. Figure 1 from Gao et al. (2013).
large values will not be visible in any of the supplementary information I am also using to improve our understanding of these objects. Thus those objects would not be included. The low resolution of the older map I am using results in the same effect.

Crowding

Another problem more likely to occur in the inner bulge region as opposed to the outer bulge and disk is crowding. Observing towards this area puts a lot of stars into the line of sight in addition to the density of the inner regions themselves.

Crowding causes problems in distinguishing stars from their neighbours - the presence of a bright star, especially one bright enough to cause diffraction spikes can "overlap" with it’s nearby neighbours, leading to inaccuracies in estimating the star’s apparent flux and colours due to VISTA’s diffraction spikes that vary with time. These inaccuracies hinder any attempts to use these stars. In dense regions, stars can overlap in the Point-Spread-Function (PSF), known as blending (Alonso-García et al., 2018). This can again result in inaccurate measurements of flux as well as confusing estimations of the background (Molnar et al., 2022). Careful PSF fitting is needed to disentangle the various stars - modelling the PSF of each close proximity object and removing their contribution from the stellar flux of the other objects (Kamann et al., 2013).

Only then can fluxes be measured as accurately as possible. VVV’s good resolution helps us deal with this issue in the near-IR, but there are no mid-IR instruments with comparable resolution and depth, leading to difficulties using longer wavelength data to supplement ours.

This requires careful source examination. This effect will come into play prominently in Section 2.2.2 and will be discussed there, with a focus on how it affect interactions with other surveys and comparing objects between these surveys. When the effect is noticeable, objects can overlap leading to measurements for objects being combined and otherwise interfering with nearby objects in the line of sight that happen to suffer overlap with them.

With the matching tools that will be described in Section 2.2.2 and overall caution with any object that differs substantially between bands or between frame measurements can help us distinguish what is genuine variability and what isn’t.
Part of the solution to the issues unrelated to crossmatching lies in the VVV PSF photometry. The default fluxes and magnitudes use aperture photometry, calculated using a 2.0 arcsecond aperture diameter with the option of changing this and using a wider or narrower one - at 2.8 arcsecond or 1.0 arcsecond diameter respectively. The tighter fit to a given object found by using PSF photometry is superior to the fixed aperture and is calibrated to a given object based on its local environment. As such, the PSF photometry is used when available.

### 2.2 VVV Observations

The telescope used for VVV, VISTA, samples the sky to create images called pawprints. With around 6 pawprints it can cover each "Tile" area, and by doing this multiple times it can ensure good coverage across the sky as well as the repeated coverage needed to judge variability (Minniti et al., 2010). As with all surveys, there is a point beyond which the faintness of the sources can no longer reliably lead to accurate measurements (as well as a point at the opposite end - where the source becomes too bright). The limits of maximum depth reachable at which we can observe are:

<table>
<thead>
<tr>
<th>Filter</th>
<th>5-sigma 5s</th>
<th>60s</th>
<th>900s</th>
<th>Sky brightness (mag/sq arcsec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>19.8</td>
<td>21.3</td>
<td>22.8</td>
<td>18.2</td>
</tr>
<tr>
<td>Y</td>
<td>19.2</td>
<td>20.6</td>
<td>22.1</td>
<td>17.2</td>
</tr>
<tr>
<td>J</td>
<td>18.8</td>
<td>20.2</td>
<td>21.7</td>
<td>16.0</td>
</tr>
<tr>
<td>H</td>
<td>17.9</td>
<td>19.3</td>
<td>20.7</td>
<td>14.1</td>
</tr>
<tr>
<td>Ks</td>
<td>16.9</td>
<td>18.3</td>
<td>19.7</td>
<td>13.0</td>
</tr>
</tbody>
</table>

Where the above table is estimated using 0.8 arcsec seeing, with a 2 arcsec diameter aperture (Minniti et al., 2010). In addition to this maximum, based on the completeness limit, there is also the reverse in VVV, the confusion limit. In Figure 2.10 it is shown how the uncertainty increases at both high and low magnitudes.

The work I have done builds upon the existing studies of variability already included as part of the VVV data. There are two main variability studies already created. The creation of the table vvvVariability was created early into the projects timeline and continues to this day (Cross et al., 2009, 2012). It contains the astrometric and photometric variability quantities drawn from the calibrated
Figure 2.10 An example of the VVV single epoch detection limits, as they are used to remove spurious detections and rejected sources selected in the mag-err plane. The darker line is fit to the logarithm of the errors at a given magnitude, whereas the grey line is the sigma-clipping limit. The darker objects are the rejected objects and the grey ones are the retained sources. Figure 2 from Mauro et al. (2013).

astrometric and photometric quantities in the detections. This includes proper motions, mean magnitudes and Root Mean Squared for the combined detections (Cross et al., 2009).

More recently, Ferreira Lopes et al. (2020) created the VVV Infrared Variability catalogue (VIVA), a larger and newer sample of candidate variables selected using correlation indices (Ferreira Lopes & Cross, 2016), non-correlated statistics (Ferreira Lopes & Cross, 2017) and period finding methods (Ferreira Lopes et al., 2020). This catalogue was based on crossmatching with other variable star catalogues and using measures derived by the authors in order to select objects with a high likelihood of being variable. The VIVA catalogue employs the New Insights into Time-Series Analysis (NITSA) (Ferreira Lopes & Cross, 2016, 2017), tailored to select and perform time-wise and light curve based analysis of a clean sample of VVV objects. It worked with the principle of dividing the data into correlated (observation occurrences with overlapping pawprints, resulting in a group of data points much more closely separated time-wise then the observations typically can allow) and uncorrelated data.

Each is handled separately, with a few different techniques proposed for each. In general, the correlated data and the associated techniques perform better at their function as indicators and are recommended to be used in place of their counterparts whenever possible. These indices measure the amount of correlation
between pairs of measurements - between all consecutive pairs of measurements over the time-series data of the variable. A more thorough description of these indices is described in Section 3.2.4.

Both of these catalogues contain very useful tools, but my work will expand upon several areas. The VIVA catalogue identifies types of variables by drawing from variable star catalogues: American Association of Variable Star Observers (AAVSO) International Variable Star Index (VSX) \cite{Watson2014} and the SIMBAD database\footnote{http://simbad.u-strasbg.fr/simbad/} - each made up of many smaller surveys with their own biases and tendencies for what they observe. If there are more Eclipsing Binary stars present in the source catalogues then this bias will be carried over to our own data. An additional factor which adds to this is the difference in observational wavelengths. The majority of variable star observations - and the majority of the cases in the sources - were all in the visible bands. If care is not taken, we risk influencing the variable stars present in our data by expecting that they represent a completely different group of objects.

\subsection*{2.2.1 Known Issues}

The science archives of the VVV contain thorough details of all known detector issues, detailed by the CASU (Cambridge University Survey Unit)\footnote{http://casu.ast.cam.ac.uk/surveys-projects/vista/technical/known-issues}. Two of the detectors, detectors 1 and 16 both have notable issues. These are large areas of bad pixels and highly variable quantum efficiency respectively. Some smaller issues in detector 4 and older observations using detector 6 also have known artifacts. These effects can be excluded relatively easily by avoiding using observations in detector 16 and relying on data flags to avoid cases where the other detectors have had issues.

There are additional issues that can arise around very bright objects. The extraction software can interpret the bright spikes around saturated objects as detections. If an object is bright enough to approach saturation the shutterless configuration of the camera produces a noticeable black spot in their centre, sometimes breaking the bright object into multiple bright spots around the black spot. The presence of this black spot is useful for easily identify such objects visually if need arises.
2.2.2 Crossmatching

By crossmatching with other variable stars surveys we can use their existing catalogues to help us understand what the variable stars we are looking for look like - especially important in the near-IR; with fewer existing time-series studies of variable stars to compare to and understand what variables we are looking at. For some types of variable star, there are likely strong similarities with their optical counterparts, but for others the matter is less clear. Hence ways to approximate and understand these differences are highly sought after.

By crossmatching with other surveys in general we can use the additional data to better constrain the objects we are working with - using other surveys with additional data and access to higher and lower wavelengths then what VVV has access to, we can get insight into the spectral energy distribution of objects and make predictions as to temperature and evolutionary stage.

When trying to study evolved lower mass stars, especially in high extinction regions, information about the spectral energy distribution - at mid infrared wavelengths can help pick out these objects - accelerating the accuracy of the machine learning process.

There are several candidate surveys which cover the galaxy bulge and disk regions overlapping the VVV observing area. The GLIMPSE and WISE surveys both extend into the mid-IR. These instruments used by each both have a larger point spread function than the VISTA telescope; WISE using PSFs with FWHM of 6.1, 6.4, 6.5 and 12 arcseconds for their four filters respectively (W1,W2,W3,W4)\(^{[1]}\) whereas the VVV PSFs are fit individually to each object\(^{[2]}\). VVV also has fixed aperture photometry with a default 2.0 arcsecond aperture. In such dense and crowded regions of the sky there is a high risk of blending and lack of clarity when matching sources together.

Since part of the crossmatching will involve a catalogue of known or suspected variables, there are certain challenges involved with doing so, as opposed to a wholesale crossmatching of different multi-wavelength data. Because we are dealing with variable objects which by definition have a range of observed magnitudes, matching must take into account that the counterpart in other surveys will also be varying and will be at unknown phase of it’s variable behavior. The matching we carry out must include more leeway than normal in order to account for this.
When comparing magnitudes between surveys it is important to understand that the filters used by astronomical surveys are not unique and to understand the difference between the filters used for each survey. Some features show up strongly inside a given passband and thus that passband is used by other surveys to focus on objects that show the same kind of feature. Whilst a given filter might have the same start and end point between surveys, its response will not always be identical \cite{Spiniello2019}. \cite{Spiniello2019} compared the overlap between the Y band data of VISTA, PanSTARRS, SkyMapper and DES and compared VISTA and 2MASS in the J, H and K bands. The greatest difference between the bands occurred at fainter magnitudes, reaching up to ±3 magnitudes in Y at the observation limit of 19 mag (again in Y band). The cause of this is more likely attributed to difficulties matching between the survey that can observe fainter objects including VVV, and the others which cannot go as deep.

If multiple objects overlap when matching, using magnitude is one way of determining which of the match candidates is the true match. As just discussed, magnitudes will differ between surveys even in the same filter. The more reliable method I found was to compare objects and their magnitudes in the context of the distribution of magnitudes between the two databases to be crossmatched.

**NWAY**

To rectify the above issues I employed the tool NWAY. NWAY is a suite of matching code and associated tools for crossmatching between surveys \cite{Salvato2017}. Whilst there already exist lists of possible matches between the VVV data and the candidate NIR surveys, NWAY provides some very potent tools to distinguish between multiple matching objects.

One of it’s key features is using magnitude (or in fact any other parameter) as part of a Bayesian analysis. By calculating the probability of an object belonging to a category of known ”good” matches compared to the belonging to the remainder of the dataset, one obtains a probability of the object fitting this selection.

The creators of NWAY demonstrate its use using magnitudes \cite{Salvato2017}. The concept of the process is noting that when matching certain types of objects to others, there may be some correlation of magnitude. An example of this is shown in Figure 2.11 using a subset of objects, in our case, variable stars.
and comparing their distribution to that of the whole sample in a probabilistic manner. Colour is another property which varies between object type, so was used as well.

I made use of part of their code and applied it to the existing crossmatching already done by the VVV VSA as part of their matching between datasets. This way would prevent having to perform the built in NWAY crossmatching on the extremely large VVV dataset. Because not all objects will have counterparts, it makes sense to enact the process of gathering the additional information to cases where it is required; relying upon the default set of information to work satisfactorily and employing the additional information as a second stage of analysis that can be performed if results of the generic method fails to work. An example of the latter case would be the densest regions where the number of potential counterparts for each star rise dramatically.
Even without running NWAY on the entire VVV catalogue, there is still a large amount of computations required that run time and computational load must be considered. In the cases where we are uncertain regarding matching an object to its counterparts, NWAY will use the object and its prospective matches as input - the distributions of magnitude and colour also required for NWAY calculated using every object in the catalogue.

If more accuracy is required, there is a more computationally expensive process possible. Instead of calculating a global histogram to relate the two datasets, the nearby points are collected and used to form a local histogram. This represents a lot of additional computation for a already slow process. As such, the global histogram of magnitude and colour is used instead. The matching counterpart assigned most likely by NWAY’s Bayesian analysis using the priors of magnitude and colour is selected as the most likely partner.

2.3 Statistical Properties

The section begins the discussion of some of the tools I have found and/or created in order to help identify, distinguish and select variable stars out of the many candidates I have - with the attention of compiling a reliable sample (> 90%) of each of the three main LPV subtypes.

One of the key tools used in study of variable objects are variability indices - properties used to calculate features associated with variability. If an object matches all the features associated with a variable object with high confidence, then we can say with confidence that the object is variable, in addition to what type of variable it is. These are mathematical values which vary in uniqueness between objects - some objects will have very similar periods, other objects will have very similar temperature, but by using both, better distinction between types can be drawn. This involves studying properties like periods, light curve statistical properties like skew and information from and about the environment and how they can best be used to extract maximum information. Examples of commonly used features not already mentioned include:

- Color
- Mean
• STD
• Skew
• Quartile Difference
• Kurtosis
• Median Absolute Deviation

These are some of the simplest ones and whilst reliable, much of the more recent work in the field of variability is deriving new and more robust features for various applications.

For this thesis, I will separate the discussion of these features into two sections. From the list above period and temperature stand out as values much more easily understood and interpreted in how they are related to the stars stage in evolution than it’s skew or quartile difference. In Section 2.3.1 there is discussion of temperature and methods for deriving it. The periodicity of these objects will be covered in Section 2.3.2 including how best to calculate the period given the data available and how reliably one can trust the periods one obtains. In Section 3.2.4 there will be a discussion of the remaining I have used, some of their derivations if needed, and the Principal Component Analysis (PCA) used to select only the ones that will give the best performance. In addition, this section will cover how they were incorporated into machine learning models to help in identifying variable stars.

Figure 2.12 demonstrates an important point regarding the application of many statistical features and variability indices on a wide group of objects with varied properties. The observations are affected by the brightness of the observing object due to how the detection instrument functions. At both the bright and faint magnitudes, there is more deviation and there are more outliers. At the bright end, saturation and detector non-linearity begins to effect the objects we observe - the non-linearity muddling the relationship between photons and electrons at the pixel and said electrons overflowing the pixel’s potential well and affecting nearby pixels. At the faint end, random noise begins to become an important factor - the noise from non-source photons from the background can outnumber the source photons and the source ceases to be detectable above the background. Even before the limits of undetectability, the added fluctuations can equal the variations in a given faint object, making it hard to distinguish the variability in
the star. The same holds true for variability indices, especially ones drawn from
the light curve itself (Ferreira Lopes & Cross, 2016).

The next section continues the discussion of some of the tools I have found and/or
created in order to help identify, distinguish and select variable stars out of the
many candidates I have. Both of these may be complicated - the potential for
additional information can help us understand many types of object, by extending
the number of observations, additional color or SED information to help get a
better picture of the object in question beyond a light curve.

2.3.1 Temperatures

Temperature information is useful information to have access to - with it,
combined with observed properties such as magnitude and colour, can give access
to other physical properties such as luminosity and stellar radius (Ramírez &
Meléndez, 2005; Takeda et al., 2002). One of the earlier plans for this involved
using this information with stellar models and with the context of stellar evolution.
tracking our stars along evolutionary tracks to give us further information such as age and mass. This aspect did not end up being a focus of this work, but the ability to estimate temperature by relying on VVV’s more extinction resistant and more numerous Ks band data still proved useful.

Even without needing to involve isochrones, the possibility of additional information - temperature and spectra - to help distinguish between something being fainter and something heavily extinguished is promising. Two options were considered for estimating temperature.

The first version attempts to fit model spectral energy distributions (SEDs) to databases of model. These model spectra databases such as X-Shooter (Verro et al. 2021) and MILES (Falcón-Barroso et al. 2011; Sharma et al. 2016) contain large grids of stellar spectra, each over a range of wavelengths. These models can be taken from real objects or artificially generated. Each model in the spectral database has known properties which enables us to, after fitting model SEDs to the observed SEDs, estimate of the stellar parameters of the observed objects by comparing which of the model spectra it best matches (Park et al. 2018). The stellar parameters depend on the model in question, but typically include effective temperature, surface gravity, metallicity, age and mass. In order to fit the model to the SED, $\chi^2$ minimization is used, where the goodness of fit, $\chi^2$ is calculated across each SED point and each model point at the same wavelength. The model with the lowest $\chi^2$ is the best fitting model, and the model’s parameters are assumed to be the closest to the candidate star.

In order to create an SED from the VVV data, more data points to make up the SED are required. The VVV survey only has 5 bands - 5 points in the SED. The more points available, both above the VVVs covered wavelength range and below it, the more accurately a model can be matched to the VVV spectra. To expand the range of wavelengths covered by a chosen object, it has been cross matched with other catalogs. These other source catalogs include Pan-STARRS, SDSS and WISE/NEOWISE, which were chosen to expand the wavelength coverage both in the redder and bluer side of the recorded VVV data. The WISE/NEOWISE catalogs are of particular use because of the heavy extinction and large reddening that occurs when looking into the dusty regions towards the galactic center.

The second version makes use of the infrared flux method (IRFM) technique (Blackwell et al. 1980; Ramírez & Meléndez 2005). This technique makes use of near IR monochromatic flux or magnitude - photometric data as opposed
Figure 2.13  (Top) Light curve of RR Lyræ variable star, phase folded, sourced from the WFCAM [Casali et al., 2007] calibration archive. Object is part of the VIVA crossmatched catalogue. Additional points on the light curve using data from Pan-STARRS (points labelled ip1,zp1 and so on), SDSS (points labelled r,g and so on) and NEOWISE (points labeled W1,W2) are coloured according to the label on the right. (Middle) Fluxes minus the average value in each band. The black point labelled roots denote points of interest such as maxima and minima. As can be seen from this and the above figure, the light curve shows the rapid peak and slow decline characteristic of a RR Lyrae, although with some scatter. The scatter is potentially due to the Blazhko effect [Szabó, 2014], which results in variation of period and amplitude over time - since this data has been collected over several years. (Bottom) The broadband SED of the above data. The spectrum peaks at around $\sim 3000K$, which does not fit the expect temperature [Catelan & Smith, 2015b].
to spectroscopic data. Existing models with known effective temperature and metallicity at a given colour are used to fit the following equation:

\[
\Theta_{eff} = a_0 + a_1 X + a_2 X^2 + a_3 X [Fe/H] + a_4 [Fe/H] + a_5 [Fe/H]^2
\]

Where \( \Theta_{eff} = 5040/T_{eff} \), \( X \) is the colour and \( a_1, a_2, a_3, a_4, a_5 \) are the coefficients of the fit. By fitting the data points to this equation, values for the coefficients can be found. With these coefficients, for a given colour pair and metallicity, the coefficients can be used to solve for the effective temperature (González Hernández & Bonifacio, 2009).

The accuracy of the technique is impressive given its simplicity. Ramírez & Meléndez (2005) report typical standard deviations on the temperature to be between 30K to 120K, González Hernández & Bonifacio (2009) find standard deviations of 100K for dwarf stars and 131K for giant stars whilst Mucciarelli & Bellazzini (2020) find uncertainties between 50K to 100K.

The latter two papers calculate the coefficients for use with 2MASS data and Gaia data respectively. The González Hernández & Bonifacio (2009) paper uses 2MASS (Skrutskie et al., 2006) data to calculate the IRFM for V-I, V-J, V-H, V-Ks, J-Ks colours. The latter one is particularly interesting for us since it can be done with no crossmatching, however it only has been calculated over a small range of colours - between 0.1 and 0.8 for both dwarf and giant stars. The colour range for V-Ks is higher - between 0.8 and 3.2.

The Mucciarelli & Bellazzini (2020) paper uses Gaia DR2 (Gaia Collaboration et al., 2016) to calculate the IRFM for BP-RP, BP-G, G-RP, BP-K, RP-K, G-K colours. These are available for a wider range of colours, with all three of the K band colour having a maximum colour of over 1.75 (Mucciarelli & Bellazzini, 2020). Using the Gaia IRFM also has the advantage of being much more available, since the VVV data has an existing crossmatch with Gaia, allowing me easy use of the Gaia magnitudes to estimate magnitudes. Mucciarelli & Bellazzini (2020) correct for reddening following the procedure of Gaia Collaboration et al. (2018c) before calculating the IRFM and as such, the VVV temperatures will be adjusted for extinction accordingly.

We may not always have access to metallicity information - relying on the Gaia
Figure 2.14  The behavior of effective temperature calculated using IRFM as a function of (BP - RP) colour for dwarf and giant stars (top and bottom respectively). The stars are grouped according to metallicity, and the different polynomials (corresponding to each metallicity) are drawn using the same grouping. Figure 1 from Mucciarelli & Bellazzini (2020).
crossmatching for it, when available. Thankfully, this is not a massive issue for the technique. As shown in Figure 2.14, the effect of metallicity does not have a huge impact on the effective temperature predicted. The effect of metallicity, at its maximum, only results in a temperature change of 200K, only slightly more that the default uncertainty. This maximum is also calculated at a very large metallicity difference of [Fe/H] = 0 to [Fe/H] = -3. This is likely to be more than any difference that will result from having to approximate metallicity, if needed.

The errors on the temperatures calculated this way must be increased slightly due to the fact that we are using slightly different bands that used by Mucciarelli & Bellazzini (2020) - Ks instead of K. The authors of the updated relation also noted the residual of their fit - $\sigma T_{eff} = 66 K$. This will be included along with any errors from the colours.

This was considered a necessary sacrifice to using this technique, and the IRFM as the temperature finding method of choice. It possesses good accuracy and only requires colour information across two bands, whereas the SED modelling required the five VVV bands at minimum, and usually required more in order to accurately match the model to the observed object. Overall, the IRFM method seems to be both faster, more accurate and is noticeably faster than scanning across a grid of spectra - since only a single calculation is required for each item.

In addition, the IRFM method is more resilient against uncertainties in extinction. Figure 2.15 shows the SED calculated under different levels of extinction. The shape of the SED is different under each, which would result in a drastically different model being fit to the same object. The IRFM method relies on the Ks magnitudes which are shown to vary much less than the shorter wavelengths, leading to it being more reliable when the extinction has large uncertainties.

An example of the performance of the IRFM is shown in Figure 2.16, which was calculated on a randomly selected sample of stars with temperatures available in the Gaia catalogue, in order to compare. The method shows that there is mostly agreement between the two samples, with a slight trend of underestimating the temperatures in the low (<4500K) and high (>7000K) temperature regions. The cause of this is not known, and given the random sample, does not appear to have any correlation to the type of object (outside of its temperature).
Figure 2.15  The SED for VVV object id 515838123764 after adjusting for low extinction, the objects extinction in our extinction map and under high extinction.
Figure 2.16  Using Gaia temperatures for comparison, the performance of the IRFM method on a random sample of stars. Both axis display the temperature in Kelvin and the black line denotes a 1 to 1 match. The scatter graph is coded based on density of points in that region.
2.3.2 Periods

As mentioned above, the period is a powerful tool to distinguish and characterise variable objects we observe (Jayasinghe et al., 2018) and as such, there has been a lot of development in different techniques of doing so. Periods can be used to select variable stars from a large survey and combined with magnitude or color data it is possible to identify regions occupied by variable types by eye (see Figure 1 in Soszyński et al. (2013b) and Figure 3 in Eyer (2006)). The section continues with the discussion of some of the tools I have found and/or created in order to help identify, distinguish and select variable stars out of the many candidates I have.

A periodogram is a method used to estimate the spectral density of a signal - in this case, the variable behavior of a star. The spectral density here describes the distribution of power to the component frequencies that make up the original signal. A frequency with a high allocation of power indicates that this frequency is a large contributor to the behavior observed. By doing this, the period(s) attributed to the variability can be deduced. The technique is an extension of Fourier methods, which are the simplest form of this (Graham et al., 2013b). By careful examination of the Fourier transform or equivalent and power spectra of a time based series it is possible to find spikes in the frequency space corresponding to periodic behavior in the original light curve data.

Periodograms are the most commonly used approach, but not the only option. Two other well-developed approaches are:

- Phase dispersion minimization (PDM) methods fold time series data as a function of phase and seek to minimize or optimize a cost function across the range of potential periods. The principle is similar to the minimization of entropy used in other problems, sometimes using entropy as the cost function for light curves, although not always (Graham et al., 2013b). This method has been recommended for objects with non-harmonic variations (Schwarzenberg-Czerny, 1989).

- String length methods use neither single points like periodograms or bins of phased objects like PDM methods, but instead use pairs or higher of objects (Ferreira Lopes et al., 2018; Graham et al., 2013b).

Some approaches combine a Bayesian approach with one/some of the techniques.
mentioned above (Wang et al., 2012; VanderPlas, 2018). A candidate is found probabilistically, not necessarily a single best fit candidate.

Since they had already been calculated, the periods already calculated by Ferreira Lopes et al. (2020) in the VIVA catalogue were considered first. In the creation of the VIVA catalogue and as part of the authors analysis they made use of the following period-finding techniques: Generalized Lomb-Scargle (LSG: Lomb 1976; Scargle 1982; Zechmeister & Kürster 2009), String Length Minimization (STR: Dworetsky 1983), Phase Dispersion Minimization method (PDM: Stellingwerf 1978; Dupuy & Hoffman 1985), and Flux Independent and L Panchromatic Period method (PK and PL: Ferreira Lopes et al., 2020).

Out of this group the most well known method is Lomb-Scargle periodogram. It occupies a position of default by dint of several reasons. It is a solidly performing baseline, used in many large surveys with good results. The mechanisms of how it works make it quite similar to the other three groups of period finding tools - it can be derived from Bayesian probability principles, for example, which gives the method and its derivatives a firm middle ground in terms of adaptability between the different groups and their advantages and disadvantages.

The Generalized Lomb-Scargle (LSG) is used over the original. It is a modified floating mean version (Zechmeister & Kürster 2009) which was found to improve period estimations for non-uniformly sampled data and help with long periods (close to the total data time span $T_{total}$). To use, this generalized Lomb-Scargle (GLS) and the original are evaluated over a range of frequencies from $1/T_{total}$ linearly down to some maximum frequency. This maximum frequency can be chosen in various ways: in some papers, the Nyquist frequency or an approximation is used (for example by Debosscher et al. (2007) allowing a slightly greater maximum frequency for variable types with very short periods); in others (Richards et al. 2011) it is chosen uniformly although with a penalty to $P_f(f)$ above the Nyquist frequency.

A formal definition and derivation of the three literature period-finding methods has been thoroughly covered in other papers (Graham et al., 2013b), which also reviews and compares period finding algorithms. The authors performed a comparison of 11 common algorithms, judging each in terms of completeness (fraction of accurate periods recovered) as a function of data aspects that might affect the choice of algorithm, such as magnitude, number of epochs, the variable objects period and light curve as well as the algorithms run time. As a function of
magnitude several effects were noted. Firstly object class affected the algorithms quite strongly, the authors noted that for test datasets with a large percentage (~50%) of semi-regular variables and other LPVs with irregular or multiple periods, overall completeness of that dataset was low due to poor fitting of accurate periods. Similarly for other datasets, the magnitude ranges which contained a large fraction (~90%) of Cepheids which were fit very well had high completeness, whereas at different magnitude ranges with lots of eclipsing binaries observed, the completeness dropped again. Because of the long baselines of the test datasets, all had high completeness for long period objects if lower accuracy requirements were needed for such objects, but fell off if more precision was needed.

In terms of the class of variable object, pulsating and eclipsing had the overall highest completeness results (~60%), with rotating objects being slightly lower (~40%). Variable categories that typically lack any kind of periodic signal such as eruptive and cataclysmic had greatly lower completeness. Similarly within each variable category, individual types of object were found at differing levels of success, for example for pulsating variables, RR Lyrae can be found at (~60%) completeness, whereas Semiregular variables were found at only (~20%) completeness.

From the completeness as a function of magnitude and period, the methods Multiharmonic analysis of variance (AOVMHW) (Schwarzenberg-Czerny, 1996), Conditional entropy (CE) (Graham et al., 2013a) and Supersmoother (SS) (Reimann, 1994) were the most successful. As a function of class, AOVMHW and CE were the most successful, unless harmonics of the true period were also included in which case LS and GLS performed best, which is believed to explain why this method performs worse at other times, due to detecting false harmonics.

The authors conclude that good periodogram algorithms could return an period with accuracy in around half of cases with strictly period objects having a higher success rate and semi-periodic or multi-periodic objects being harder to fit a period for. The situation is further complicated by some periodic variable objects changing period over time. Several other methods are suggested, but all also suffer downsides that prevent them being a replacement primary period finding method instead of periodograms. This lack of a guaranteed find is unfortunate, but this does not prevent attempting classification and further analysis, as described in Chapter 3.
For some variable stars including Cepheids, RR Lyrae and LPVs, the variations can be multi-periodic or oscillate in different modes. In order to find and identify these objects, at least some of the periodogram methods used must be able to deal with this eventuality. Following Debosscher et al. (2007) multiple periods can be fit by fitting a light curve with a linear term plus a harmonic sum of sinusoids. Each sinusoids frequency is first found from the initial periodogram, and then that peak, plus additional harmonic frequencies at multiples of this. After removing this signal from the data, a periodogram is fit to the remaining data and if a good enough fit is made, the process repeated until all strong peaks are fit. The amplitudes and phases of the frequencies and harmonics are noted, as they can also be useful in classification.

The last two periodogram methods were created by the VIVA authors (Ferreira Lopes et al. 2018):

- Period algorithm PK - based on the even dispersion, flux independent period indices “kFi” designed in Ferreira Lopes & Cross (2016) and implemented as part of a period finding process in Ferreira Lopes et al. (2021).

- Period algorithm PL - based on the even dispersion panchromatic measure “L” designed in Ferreira Lopes & Cross (2016) and implemented in Ferreira Lopes et al. (2021).

Both methods have the same underlying structure. In order to find the best fitting period, the time-series data folded with a given frequency and the scatter in phase space is minimized across all tested frequencies. The frequency that produces the smallest scatter is taken as the estimate of the period. Where the two approaches differ is in how they characterize the scatter in phase space (Ferreira Lopes et al. 2021). Both approaches combine the data from the multiple bands of VVV and measure the scatter using the correlation of phase but calculate the end measure slightly differently as shown in equations 2.2 and 2.3.

\[
P^K_S = \frac{N^+}{N}
\]  

(2.2)
\[ PL^s = \frac{1}{N} \sum_{i=1}^{N} Q^s_i \]  

(2.3)

These equations are defined for a given number of correlated observations \( s \) and are calculated using \( Q^s_i \) as defined in equation 2.4 \( \Lambda^s_i \) in equation 2.5 and \( N^+ \) means the total number of positive correlations drawn from equation 2.5 measuring objects positively, negatively or neutrally aligned with their neighbours.

\[ Q^s_i = \Lambda^s_{i,i+s-1} \sqrt{|\delta_i \ldots \delta_{i+s-1}|} \]  

(2.4)

\[ \Lambda^s_i = \begin{cases} +1 & \delta_i > 0 \ldots \delta_{i+s-1} > 0 \\ +1 & \delta_i < 0 \ldots \delta_{i+s-1} < 0 \\ 0 & \delta_i = 0 \ldots \delta_{i+s-1} = 0 \\ -1 & \text{otherwise} \end{cases} \]  

(2.5)

\[ \delta_i = \sqrt{\frac{n_w}{n_w - 1} \frac{y_{i,w} - \bar{y}_w}{\sigma_{i,w}}} \]  

(2.6)

where \( n_w \) is the number of measurements, \( y_{i,w} \) are the flux measurements, \( \bar{y}_w \) is the even-mean computed using those observations (a variation of the mean derived in Ferreira Lopes & Cross (2017)) and \( \sigma_{i,w} \) is the flux error of a given waveband \( w \).

The period finding for these 5 methods was carried by Ferreira Lopes et al. (2018) as part of the creation of the VIVA catalogue out on the VVV DR4 data over a frequency range of \( f_{\text{min}} = 2/T_{\text{tot}} \) to \( f_{\text{max}} = 30 \) and a frequency sampling of \( N_{\text{freq.}} = 20 \times f_{\text{max}} \times T_{\text{tot}} \) were used where \( T_{\text{tot}} \) is the total baseline of the observations.

The periods found were compared to literature periods to test the effectiveness
Crossmatched periods - labeled literature periods here - against the periods calculated using the 5 different methods used by VIVA. The periods are plotted in log10 base, and the scatter plot is coloured according to its density. The objects used here are the subsample VVV-CVSC-CROS - defined as objects with matches in the AAVSO or SIMBAD database.
of the method, which is shown by comparing the objects with known periods in the VIVA catalogue with their periods found using each method in Figure 2.17.

In overall performance, the LSG and PDM methods return the highest yield, where the yield is defined as the fraction of objects inside the first harmonics - direct multiples of the true period of an object. The lowest harmonic is visible in all of the plots in Figure 2.17 as the second straight diagonal line below the \( x = y \) line that corresponds to a one for one match. This lower harmonic corresponds to half the true period, with the high harmonic correspond to double the true period. The yield for the LSG and PDM is 73% and 71% respectively, with the other methods being STR: 59%, PL: 59% PK: 42%. In the release of the VIVA Catalogue (Ferreira Lopes et al., 2020), the candidate objects included periods found using multiple period finding approaches. This include best fitting period, calculated by comparing the signal-to-noise value for each method.

There are noticeable horizontal lines correspond to periodic behavior in the observations. These lines correspond to multiples of the sidereal day, solar day and synodic month cycles (Dawson & Fabrycky, 2010). These over-abundances at specific frequencies can also be caused by aliasing (Ofek et al., 2020).

From the objects in VIVA with known counterparts with existing periods the following types of variable produced periods with more disagreement with literature that others. These were RR Lyrae (~12%), eclipsing binaries (~12%) and SRVs (~55%). The disagreement with SRVs is somewhat to be expected. These are objects that can show multiple periods and have the noisiest and most irregular light curves out of the LPVs (Soszyński, 2022).

A downside with these VIVA periods is that the data used to calculate them, VVV DR4, has fewer epochs compared to other large scale variability surveys. Generally, one desires 100 data points in order to get reliable periods (Graham et al., 2013b), but the lack of data points for each object means that period finding methods will be hampered. Combined with the performance of the existing period finding techniques on Long Period Variables has been demonstrated to be sub-par - this is a hindrance that must be addressed. The different types of Red Giant and AGB stars can face different difficulties finding their true periodicities.

Some variability studies (Guo et al., 2022) must also account for changes in variability between cycles. The most common examples of this are changes in period or amplitude, often observed in RR Lyrae. Due to the longer base periods of the LPVs I am using, this becomes less of a prevalent issue as there will be
fewer cycles for the LPV to change over, compared to a short period object. Nesci et al. (2022) have found no significant period change over 50 years of a sample of AGB LPVs.

2.3.3 Improving Period Finding

There are several ways to improve our period finding and better identify if an object has an incorrect periods.

A simple sounding idea is to use the periods calculated by other sources for these variable objects where available, but this is not reliable. Iwanek et al. (2021a) demonstrate how properties of variables objects can vary between different observing wavelengths. They demonstrate phase-lag for LPVs, reaching ± 0.2 out of phase between optical and near-IR.

We can use other approaches to help classify objects - a Mira identified with high confidence will allow rejection of spurious low periods - and period finding can be done so as to reject erroneous values. Knowing existing information about an object with confidence can be used to weight periods in the more likely range higher. Although they will have periods closer to what the period finding process seems to trend towards, identifying them in the first place is difficult. In terms of amplitude and period they closely resemble Eclipsing Binary stars, especially when there is a lack of data points covering the eclipse.

One of the biggest and simplest ways to improve upon the previous results is to increase the amount of data points available for period finding. With the recent release of the VVV DR5 data, I now have access to many more epochs of data, with the typical light curve amount increasing from between 50-100 on average to between 100-150 on average. This new data comes from the switch to including all paw-print observations and the release of new data. This can be used to improve period finding accuracy, as the accuracy of all methods scales with the number of data points.

To calculate the new periods using the improved DR5 data, two periodograms were used:

The first method is a default Lomb-Scargle algorithm, the results of which are demonstrated in Figure 2.18. As expected, the highest spike has noticeably
Figure 2.18  Crossmatched periods - labeled literature periods here - against the 4 highest periodogram spikes for the Lomb-Scargle algorithm. The first graph corresponds to the highest periodogram spike, the second corresponds to the second highest periodogram spike and so on. The periods are plotted in log10 base, and the scatter plot is coloured according to its density.
cleaner relation, although areas overlapping the SRVs and RR Lyraes in the VIVA period still exist in this version. The curves in graphs 2,3 and 4 are believed to be due to aliasing from the analysis process (Ofek et al., 2020).

The second method is Cauchy-Schwarz Quadratic Mutual Information (QMI) minimization (Huijse et al., 2018), the results of which are shown in Figure 2.19. This method functions by epoch folding across a range of trial periods to obtain a phase diagram. By calculating and maximising the Mutual Information between phases and magnitudes - maximizing the amount of information carried by the observations (Huijse et al., 2018). The method is implemented by the python package P4J. Using the P4J package, I created code to find improved periods for all of the VIVA candidate objects.

Comparing the two methods, both methods show artifacting lines corresponding to systematic patterns in the data. The QMI has noticeable horizontal line at the

https://github.com/phuijse/P4J
one day mark which corresponding to the instrument observing each night and the LSG algorithm shows the same structure but less strongly. Similarly, there is a line at several hundred days in the QMI. The noticeable lines for the LSG periods are vertical instead of horizontal but I am uncertain if they are literature objects assigned an artifacted period by their original author or truly long period objects falsely labelled. The latter seems less likely as the objects in question were classified by the LSG as a wide range of periods - if they were long period objects falsely labelled, I would expect to see their periods a mix between their true long period and easily identifiable artifacting lines.

2.4 OSARG Specific Issues

After crossmatching the VVV catalogue with OGLE-III’s catalogue of Long Period Variables (following Section 2.2.2) to acquire VVV counterparts to preexisting LPVs, I noticed several concerning issues.

![Figure 2.20](image)

**Figure 2.20** Mean magnitude across the matched Ogle OSARGs and their VVV counterparts. KsAperMag3 denotes the VVV Ks band magnitude with a 2.0 arcsecond aperture.

The first issue was that, under the default matching parameters, the OSARGs
were being matched to extremely bright stars - with K band magnitudes of up to 10. At this brightness, this puts many of the objects at risk of saturation. Whilst the OSARGs in the OGLE sample were indeed bright (∼12 Mag in I and 15 Mag in V), it was not expected that so many would have saturated counterparts in VVV. This was observed even in the Bulge samples, likely to contain the more distant objects. The concern here was that if the OSARGs were brighter in our observations by more than I expected, it would be uncertain if there would still be enough to work with at the fainter end. An example of the matching magnitudes are shown in Figure 2.20. There is clear bimodality in the histogram of the Ks magnitudes, but the cause of the spike is unknown.

The OSARGs are being matched to very bright objects, with a strong difference in magnitudes between the OGLE measurements and the VVV ones. As shown in the mag histogram before, with the VVV counterparts of these objects being much brighter. This is expected for OSARGs - as red giants they are brighter in VVV’s filters and are dampened less by extinction. For comparison, the OSARGs in Figure 2.20 are contrasted with Mira in Figure 2.21 and Cepheids in Figure 2.22. Both groups of VVV Ks magnitudes are very sharp and both are congregated at brighter magnitudes. When comparing the presence of each spike, each one seems to correspond to one of the bimodal humps seen in Figure 2.20.

The second issue was also noted here. The OSARGs are so named because they have small amplitudes. As shown by the light curves in Figures 2.23, 2.24 and 2.25 and the histogram of amplitudes in Figure 2.26, all of the OSARGs have the small amplitudes expected in the I band data, however in the Ks band used by the VVV, the amplitudes seen are ∼5x greater, potentially even 10x greater. This is most puzzling because LPVs, according to literature, should have smaller amplitudes at longer wavelengths - with Miras decreasing in amplitude from highs of 8 mag to a max of maybe 2 or 3 in near-IR (Iwanek et al., 2021a), although there does not seem to be an agreed-upon physical explanation. This leaves the high amplitude of the OSARGs as a puzzling conundrum.

Examples of amplitude comparison are repeated for Mira in Figure 2.27 and Cepheids in Figure 2.28. The Mira have lower amplitudes in Ks-band than in I-band, as expected, with slightly lower amplitudes in Ks-band over I-band for the Cepheids. In both cases there is slightly more overlap than I was initially expecting and both peak at higher amplitudes than predicted. Despite these two points, the overall trend is the same: the amplitudes are greater in I-band. This is the exact opposite of what occurs for the OSARGs.
Several theories and hypothesis were postulated:

**Peculiarities of the VVV survey?**

One explanation I put forward is that because VVV is a variability survey, designed with cadence and depth for Galactic plane variables such as Cepheids and binaries, that these design choices make it poorly suited to detecting the OSARGs tiny variations in the near-IR at all, and all the objects in my samples are being matched to something else. The hypothesis was that other objects being falsely matched are what is creating the large amplitude changes, not the OSARGs. The argument against this asks if this were true, why are the matching objects still showing large amplitude variability? One would suspect that an average object would not vary - showing only random noise around a constant level. This is not observed in almost all cases. A simple explanation exists for the very bright cases - these objects are saturated and the large jumps are due to bad saturated measurements. But if this is the case, why is occurring for objects at ~14 mag and fainter in Ks? This is healthily below the point where saturation should begin to be an issue ([Smith et al.] 2018), at around 10-11 mag in Ks.
Figure 2.22  Mean amplitude of OGLE Bulge Cepheids and VVV counterparts. The vertical axis denotes the relative density in each histogram bin.

It was considered a possibility that perhaps this bright sample might be saturated earlier than expected - thus for a sample of these objects, I have checked the pawprint images themselves to examine by hand for any sign of saturation - diffraction spikes and holes in the centres of images and examined the RA and DEC measured for greater than average variation from observation to observation, caused by an object not being resolved properly. An example of a saturated object near to a non saturated object is shown in Figure 2.29, with notable distinction between the saturated and non-saturated. Whilst some of the objects matched with OSARGs did show some signs of saturation, the anomalous behavior still occurs even when such objects are discarded. Likewise, these objects showed no positional uncertainty above regular levels - as some saturated objects can jump in sky position due to saturation causing bright spikes and dark spots at the centres of objects for well saturated images - causing the measured centre of the light from an object to not always been in the same place.
Another possibility is that the dense nature of the bulge combined with VVV’s improved resolution over OGLE gives VVV many more observable objects, potentially leading OGLE objects having counterparts that the original survey didn’t observe.

To test this, the outer regions of the bulge were probed. When focusing on these regions more distant from the galactic centre \(|b| > 4\), the effect still seemed to be present, suggesting that it is not due to overlap and stars not getting matched properly. For comparison with this, Alonso-García et al. (2018) compared regions of the sky in the inner bulge, outer bulge and the galactic disk. They observed drastically (> 10%) lower completeness at lower magnitudes, reaching (50%) lower at fainter magnitudes in the inner bulge than the other two regions. The disk still features strong extinction, but less crowding, further confirming that crowding is the key issue at play. Unfortunately the initial detection of OSARGs in the OGLE III data was only carried out in the bulge region, preventing checking for OSARGs in the disk. Even so, at such high galactic latitudes there should be
some noticeable decrease if overcrowding was the problem.

**Systematic Issues**

The possibility of systematic issues within the VVV observation or processing stages is also considered.

There are some known issues with the VISTA data. Some of the detectors have known issues, including detector 16 (extNum=17 in the data schema) with quantum efficiency high enough that flatfielding becomes impossible in some parts of the detector. Additionally, the amount affected is variable in time. A full description is contained within the Cambridge Astronomy Survey Unit (CASU) [4]

When tested, removing all points taken with the faulty detector did not dampen the effect. An examination of the data grouped by detector used in observation also failed to reveal any clear differences between said detectors with regards

to these larger amplitudes. It is beginning to be apparent that the behavior is affecting much of the collection of OSARGs.

This phenomenon is also only occurring in OSARGs specifically. Identical Cepheids do not have the same magnitude in OGLE and VVV data, but are the same scale and obey the predicted relationship of variability being greater at shorter wavelengths that the rest of the pulsating variables obey.

2.4.1 Effects at other wavelengths

Further crossmatching was performed to expand the coverage of the data available to me. With OGLE data covering the wavelength bands V and I and the VVV covering the Ks band (with multiple epochs), I employed PanSTARRS (Chambers et al., 2016) data to bridge the gap between the two systems to examine if there was a recognizable wavelength-dependent trend.

There is filter overlap between PanStarrs, OGLE, the Zwicky Transient Facility (ZTF) (Bellm et al., 2019) and VVV but difficulties arise when trying to use

Figure 2.25  Light curve of OSARG, showing both VVV and OSARG data for a matched object. Object ID in VVV DR5: 515847254537
Figure 2.26 Mean amplitude of OGLE OSARGs and VVV counterparts. The y axis describes the density in each histogram bin and the x axis shows the amplitude in each bin for the VVV Ks band and the OGLE I band. VVV objects with magnitudes approaching saturation threshold (<13 mag) were removed, using a total of 147585 objects.

PanSTARRS. Their observations include multiple observations in each of the 5 main bands they use (g,r,i,z,y), which means that each individual band by itself only has (typically) a fifth of the total observations.

It is not possible to simply combine bands to increase the number of observations - as demonstrated by the filter plots, there are substantial difference in wavelength between each band.

An example comparison of PanSTARRS light curves is shown in [2.31]. The data, whilst having multiple epochs in each filter band, lacks a single band with a large number of points. The lack of light curves with evidence of cycles or other way to reasonably draw an amplitude from hinders my ability to obtain a usable amplitude.
2.4.2 OSARG Simulations

The discrepancy between the OGLE OSARG amplitudes and the amplitudes of their VVV matches does not seem to be explained by instrumental or the effects of clustering and density or bad matches between objects. The last possibility is that the difference is due to a process going on in the star itself, an interesting possibility if true.

Similar phenomenon (decreased amplitudes) and previously been observed for Mira variables [Smith et al., 2002; Whitelock et al., 2006]. New models of circumstellar dust shells around Mira variables was the suspected cause. In addition to amplitude variations, there was evidence of a time lag between the optical and IR observations of the Miras - where the light curve was behind the maxima of the optical pulsation cycle by $\sim 0.05-0.1$ [Smith et al., 2002]. Curiously for some objects, the mid-IR maxima occurred before the near-IR maxima. No research into other LPVs was found, however.

To test if difference is due to an internal process would require spectra of the star at different points throughout it’s pulsation period, which would indicate the
amplitude of the pulsation across different wavelengths. Similar results have been discovered for Mira variables in the visible and NIR, but OSARGs are relatively new.

Large surveys with spectroscopic data were examined, primarily the recent SDSS DR17 release, but the main OSARG sample lies in the Galactic Bulge with no counterpart spectra available in the Sloan data including the recently released SDSS DR17 (Abdurro’uf et al. 2021). Similar issues occurred with Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) (Cui et al. 2012) and it was resolved that the models would be the only reasonable way of testing this potential effect.

At <6 months from the submission deadline to this PhD, there is no time to submit and carry out an observing proposal for any new spectroscopic observations of OSARGs. I must make do with models of red giant stars and make approximations of the objects in question.

Like any other pulsating variable, the OSARG physically expands and contracts over the course of it’s cycle, resulting in different luminosity, temperature and
The physical size of the star.

To simulate this, comparing adjacent model spectra - with similar initial conditions and differing temperature and surface gravity, we end up with a small set of objects with the same underlying composition but with different sizes and surface temperatures.

The research I have found modelling Mira suggests that the surface temperature can change by as much as \( \sim 500 \text{K} \) for a Mira of 3227 K mean temperature \([\text{Iwanek et al.}, 2021a]\). With the physical changes being on such a scale, I believe the above modeling process will not deviate drastically from the truth.

The spectral library used for this is the X-shooter DR2 library \([\text{Verro et al.}, 2021, \text{Gonneau et al.}, 2020]\). It has a sample of \( \sim 800 \) spectra of \( \sim 660 \) stars, collected by the X-shooter spectrograph and consisting of separate three spectra segments observed simultaneously, giving wide wavelength coverage, reaching from the end of the UV to the start of the NIR. Their studies contain matches with Vizier and temperature calculations - most objects have a predicted spectral type, some are labelled as LPVs and most have temperature calculated separately with the UV and visible data. The ranges of [Fe/H], logg and Teff were calculated by \([\text{Gonneau et al.}, 2017] \) and \([\text{Arentsen et al.}, 2019]\) and shown in \(2.32\).

To examine how the light from the object would appear in the different filters our instruments used, each spectra was convolved with the following filters: V,I,Z,Y,H,J,Ks; Representing the coverage I would have combining the data.
of the OGLE and VVV surveys. As above, the filter information was acquired from the SVO filter profile service (Rodrigo et al., 2012; Rodrigo & Solano, 2020). The python package Astropy (Astropy Collaboration et al., 2013, 2018) contains a convolutional filter package - Astropy.Convolution - which was used to perform the convolution.

To obtain as accurate simulations as possible it is important to consider the factors that would stay the same during the variation cycle and those that would differ. I will assume the mass and composition of the star will remain constant - I am working with evolved objects they may well be losing mass and blowing off their outer layers as stellar wind, but our observations are occurring on a drastically faster timescale than the evolution of these fairly low mass objects. Most OSARGs are red giant branch and horizontal branch stars so I believe it is a relatively safe assumption to treat objects in the model library with comparable metallicities as good approximations of objects with similar formation.

Literature typically describes the mass-loss of RGB stars as having minor impact on the RGB phase but impactful on later stages of evolution. It is noted as being notoriously difficult to measure and papers prefer to describe the mass loss between points in its evolution rather than as $M_\odot \text{yr}^{-1}$ (McDonald & Zijlstra, 2015). A paper by Mullan & MacDonald (2019) convert that measurement into a prediction of $4 \times 10^{-9} M_\odot \text{yr}^{-1} \sim 2 \times 10^{-7} M_\odot \text{yr}^{-1}$, which rough overall mass loss of a quarter of a solar mass.

A similar approximation cannot be done with mass, the catalogue does not include such information. Only the likely spectral type and observable physical parameters are available.

The process begins with comparing objects of the same (or as close as possible) metallicity and slightly different effective temperature and gravity. This was done in python. The filter transcription curves were sourced from the Virtual Observatory filter service. The following process was used:

- Normalize spectra - the difference being tested for is a variation between a single object’s different states. It is expected that the spectra will change shape but the amplitude of such variation is already known. Large differences are due to differently generated spectra.

- Trim the spectra - the models contain areas affected by Slit Loss Correction. An example of why this is necessary is shown in 2.33

90
• Integrate over the spectra - For each applicable filter, convolve the filter with the spectra. The integral represents the approximate intensity of the object in the wavebands our instruments can detect \cite{Tokunaga & Vacca 2005} and is given in equation \ref{eq:2.7} where $R_\lambda$ describes the wavelength dependence of the filter response function and the spectra in question denoted $f_\lambda$. The resulting flux is calculated by converting to frequency space and then to AB magnitudes, as described in equations \ref{eq:2.8} and \ref{eq:2.9}.

$$\langle f_\lambda \rangle = \frac{\int R_\lambda \lambda f_\lambda d\lambda}{\int R_\lambda \lambda d\lambda} \tag{2.7}$$

$$f_\nu = \frac{\lambda_{\text{eff}}}{c^2} f_\lambda \tag{2.8}$$

$$m_{\text{mag AB}} = -2.5\log(f_\nu) - 48.60 \tag{2.9}$$

Figure \ref{fig:2.33} reveals the similarities in spectral shape between both a cool evolved object and subgiant. The distinct noisy patches do not have much overlap with any of the filters I am using. There is a noticeable spike in spectra where it overlaps with the Ks band filter. If this is an artifact, like the areas affected by Slit Loss Correction, it is far smaller in scale which makes identifying it unclear.

Overall, the X-Shooter models are fairly uniform and featureless where they overlap with the Ks band and between models of close temperature there is very little change in transmission where convolved with the Ks band. The I band shows more change in transmission between spectra but no visually obvious dips that might correspond to the I-band being dampened. I would have expected the greater variance in transmission at I band to signify a greater change as the OSARG goes through its cycle, but again, this does not occur. I do not have time
remaining to undergo a full spectral study using these models so the OSARGs large amplitudes remain inconclusive, but I believe that this approach is still worth considering, especially if additional spectral information about OSARGs can be gathered which would also improve the spectral models available.

Regardless of the questions still to be answered, useful temperature and periodicity information about the variable stars in question will be an asset for the classification task.
Figure 2.30 Filter efficiency curves for the filters used in this project. The plots take the form of filter transmission against wavelength. Filter data provided by the SVO filter profile service (Rodrigo et al., 2012; Rodrigo & Solano, 2020).
Figure 2.31  Light curves of a sample OSARG using panSTARRS data. The additional points present in the $z$ and $y$ filters are the mean values of the VVV data in these bands, in the VVV counterparts of those filters. Object ID in VVV DR5: 515909698212
Figure 2.32  Atmospheric parameters of the 769 stars in the X-Shooter library. Star shapes denote slightly older values by Gonneau et al. (2017) and circles are more recent values from the first data release by Arentsen et al. (2019).
Figure 2.33 Two X-shooter models normalized to the higher amplitude spectra, with the filters overlaid for comparison of wavelength position. The orange coloured object is X-Shooter object 374, HD 134439 and the blue coloured object is X-Shooter object 300, HD 170756. HD 170756 is believed to be a post-AGB star with temperature around $\sim 5300$ K and HD 134439 is believed to be a subgiant star with temperature around $\sim 5100$ K. Both have similar metallicity. The noticeable "blurry" regions at 17000 A and 13000 A are examples of regions affected by Slit Loss Correction and are the two biggest regions affected. There is also a similar area around 5000 A, which overlaps the OGLE V band filter.
Chapter 3

Machine learning for variable star classification

3.1 Introduction

In large surveys such as VVV, the rapid growth of object collection may outpace the ability of the researcher to process said data. Large teams are required to ensure maximal extraction of science information from the data. Even so, outlier cases still go under the radar. For example, in the upcoming Transiting Exoplanet Survey Satellite (TESS) mission (Ricker et al., 2015), they have an input catalogue of \( \sim 470 \) million sources with \( \sim 1000 \) epochs for each (Stassun et al., 2018). Automation is required in order to be able to cope with such data flow.

As information increases, our understanding of what is relevant in that information must also increase, if we are to continue to be able to effectively study it. Sections 3.2.4 and 3.2.4 will describe variability indices - derived tools that let us make broad cuts and selections to our data as well as how we know which ones matter. These tools do not solve the problem on their own - automated techniques capable of pattern recognition can make use of our improving knowledge of what constitutes good data (Baron, 2019) combined with the computational speed on modern machines (Sen et al., 2022) are used more and more (Bai et al., 2018).

One powerful tool for pattern recognition and appropriately studying large scale patterns in the data is machine learning. At its core, machine learning is focused
on the concept of a computer program capable, to some degree, of self adaptation. This facet of change without the need for a human programmer to manually adapt the program for every combination of input and output, makes this technique sound very appealing in times where there is more data than ever for the data scientist to deal with. An extension of this concept is that if a network is capable of learning and picking up the patterns that humans have observed, then they can detect patterns that we weren’t able to observe and grow our understanding from this.

The potency and effectiveness of machine learning is demonstrated by the many surveys (Garcia-Dias et al. 2018) incorporating the technique into their analysis (Bai et al. 2018), or even into their data pipeline processes (Mahabal et al. 2019).

The goal of this chapter is to detail two separate methods of training a network to be able to tell variables stars apart. The basic process involves selecting what input the network will study, splitting the data into multiple sets, some for training the classifier and some for testing it, tuning the hyperparameters of the model (tuning the underlying properties of the model) (Sesar et al. 2017), training the model on the training data, and then testing its performance on the testing data. Not splitting up the data like this leads to overfitting – where the network becomes very good at recognizing and classifying the input data but loses all flexibility and ability to handle or understand new inputs. The training data is used for training the network and the testing data is used for evaluating the trained network, with no overlap between the two groups.

The machine learning methods used in this thesis are examples of supervised machine learning - where the network learns from pre-existing data - labeled with known information found beforehand, learning to map the features of this data to output labels that denote what the object is (Mahabal et al. 2017). When trained, the network is able to classify new data it receives based on the input it received.

What these features are varies between the network, its design and the data it works with. Some networks can better work with more features than others or some data could be processed to extract features that contain a lot of information that helps the network classify them. Random Forests (Breiman 2001a) are a classifier with a lot of use (Plewa 2018; Carliles et al. 2010) which makes use of variability indices (Sesar et al. 2017), as mentioned in Section 3.2.4, as well as period, Fourier magnitudes (from Fourier period fitting), and colour information.
There are many other types of classifier outside of random forest. Elorrieta et al. (2016) and Hastie et al. (2009) provide examples and descriptions of several. They reported that the methods of stochastic boosting and AdaBoost (adaptive boosting) (Freund & Schapire, 1996) performed best out of nine different classifiers based on area under curve, precision, recall and F-score. Stochastic boosting works by randomly selecting a subsample of the data at each iteration, improving the flexibility of the network (Friedman, 2002). The AdaBoost algorithm works by weighting difficult to classify instances more highly and then training new branches to focus on these more difficult patterns (Freund & Schapire, 1996).

Random Forest was roughly third, performing slightly worse than multiple hidden neural networks (MHNN) and deep neural networks in terms of recall and overall F-score, but did better in terms of precision. Ease of use is another consideration in choosing a machine learning model. The models used are defined in Sections 3.3 and 3.4.

Classifiers are not without their limitations. The training data should be suitable for the task at hand. If the training data covers regions of the sky different to what the classifier will be used on, then differences in metallicity in the regions, age and reddening can hinder it’s accuracy. Similarly, if the classifier is not sufficiently trained on variables of a specific type, it may struggle to recognize them.

### 3.2 Selection of Candidates

#### 3.2.1 Training Data

For the training process, known variables are required. The VIVA Catalogue contains ~40 million variable star candidates (Ferreira Lopes et al., 2020), drawn from ~280 million sources. The VIVA Catalogue’s high completeness leads to a predicted 90% contaminant ratio - ending with a final predicted variable star collection of ~4 million objects. Roughly 340,000 of the candidates were successfully crossmatched to the variable star collections of SIMBAD [1] and the American Association of Variable Star Observers (AAVSO) International Variable Star Catalogue.

Star Index (Watson et al., 2014). This catalogue forms the primary initial selection of candidate variables. An important consideration about SIMBAD and the Variable Star Index are that they are conglomerates created from many separate studies of stars. This means that they are not always internally consistent with how objects within are found or identified.

The crossmatching is a useful bonus but is somewhat limited by the number of LPVs from the crossmatched catalogues. Between the AAVSO and SIMBAD databases, over a third are labelled as eclipsing binary (~170,000). In previous versions of this crossmatching (private communication with Ferreira Lopes), the eclipsing binary fraction - specifically EA-type binaries (Algol Type Binaries) were over ~90% of the total matched variables. This fraction was most likely erroneous - EA type binaries in the AAVSO number at 95965 compared to 392675 EW type binaries and 338166 unsorted eclipsing binaries (Watson et al., 2014), but demonstrates the sheer volume of eclipsing binary variables in the AAVSO and SIMBAD source databases.

To adapt to the earlier version of the catalogue with a heavy bias towards one particular category of object meant dealing with concerns of overfitting and training issues - if EA variables numerically dominated the training data, the trained machine learning networks run the risk of learning the wrong lesson. If the classes of object used to train are dominated by one type, the network can learn that it achieves satisfactory results by predicting that every item belongs to the dominant group (Graff et al., 2014).

For other categories of stars, the well classified Mira-type variables numbered at a ~2,500 and the far less defined category of SRV tally at ~71,000. It is the authors opinion that SRVs may be labelled as such due to inconclusive evidence. For example a labelled SRV may be a Mira that lacked sufficient coverage and enough epochs to accurately determine it’s period and only shorter secondary periods were observed and as such was labelled falsely as semi-regular - SRVs are a clearly defined group. Because of the composite nature of the AAVSO and SIMBAD databases, stars within are not labelled consistently or using the same standards. Because of these two compounding factors, it was considered prudent to only use SRV that demonstrated a good coherent period.

It is worth mentioning again the problem with overfitting, given how one category of stars will be selected far more stringently than others. This would normally lead to an unbalanced set of training data, but SRV are much more common than
Miras, as described above, so the numbers of the two are unlikely to cause an issue, especially when compared to the number of OSARG variables. A second, less obvious issue is that for Miras and SRVs compared to OSARGs, the longer period objects will have more observations on average simply because the base amount of observations required in order to measure their periods in the first place is higher. Using any property based on the number of observations could teach the machine learning algorithm that more observations means a longer period object, which is not true. Such a property will be avoided in the teaching of the network.

The OGLE surveys released up to this point revealed $\sim 350,000$ LPVs (Soszyński et al., 2009, 2011a, 2013a), supplementing the initial sample in VIVA by a factor of five. The majority of these objects are OSARGs ($\sim 300,000$), with $\sim 40,000$ SRV stars and $\sim 10,000$ Mira variables. Using the newest OGLE IV data release, Iwanek et al. (2022a) have increased the number of known Miras to $\sim 65,000$, with likely many more OSARGs and SRV to come. These objects - as well as any counterparts found by matching the OGLE objects to other surveys form much of the test and training data. This means we must be aware of the observational properties of the OGLE data. Because VVV is able to see much deeper than OGLE, this suggests that the complete sample of LPVs observed should be of this order of magnitude or greater.

The sources have different amounts of information available. The VIVA catalogues have many statistical measures (described in Section 3.2.4) which do not exist for the OGLE data, and sometimes missing for the VVV counterparts of said OGLE objects. The differences between the two make combining the two sources into one combined dataset more difficult and lead to the decision to handle the two mostly separately - using the OGLE light curves for the light curve based convolutional neural network (CNN) in Section 3.4 and the statistical features of VIVA for the statistical features based random forest in Section 3.3. Using simpler statistics such as skewness and amplitude (also described in Section 3.2.4) could be used in place of the statistical measures used in VIVA and were considered in the same section. The eventual choice of features, in this case no pre-calculated features, is described in Section 3.4.

In addition to OGLE data, I make use of Gaia EDR3 data (Gaia Collaboration et al., 2016; Lindegren et al., 2021) - crossmatched with VVV for 2 reasons. The first reason is for access to Gaia colours to estimate temperature following Section 2.3.1. The second is for access to Gaia parallaxes - if any period-luminosity
relation is to be established, it requires objects with known distances to calibrate it.

Although not particularly a problem here, techniques have been created to expand training sets in a process commonly called augmentation. Additional pieces of training data are created by creating variants of the real objects. For networks based on imaging, flipping and rotating real images both provides the network more objects to learn from but also trains it to be capable of recognizing an object regardless of real or artificial perturbation (Bowles et al., 2018). Similarly, the same principal can be applied to light curves. As light curves have a fixed shape and cannot be flipped or inverted, generating new light curves (if the source is well understood) or creating variants of real light curves by adding noise or sampling them - creating a non-identical copy that still captures the variability but with slightly less accuracy (Bowles et al., 2018). In a similar vein, when creating the training set it can be extended past the amount of data available to increase the classifiers ability to handle certain phenomenon, such as reddening. By adding additional de-reddened objects with magnitudes adjusted to have similar brightness to the rest of the training data after de-reddening occurs, the classifier can better learn how to deal with different amounts of reddening (Hernitschek et al., 2016).

### 3.2.2 NIR and Optical Properties

The section will examine the NIR properties of the subclasses found using the previous chapter and previous section. There are few examples of large numbers of objects studied at these wavelengths, this section will provide new information before examining properties like how the spectral energy distributions (SEDs) of variable stars change with phase. This will be helped by compounding the NIR data with available optical and MIR data where possible.

For the purposes of variability studies like this project, there are challenges associated with operating at these wavelengths as well. Fewer studies at these longer wavelengths leaves less data to draw conclusions about Period-Luminosity relations from, leaving many unknowns. There is still lots to be learned from near-IR surveys, as they are still quite new - with near-IR detectors being ~ 20x as expensive due to greater manufacturing complexity and smaller market (Sutherland et al., 2015). However this also means that many variable stars have been studied relatively little at longer wavelengths (Angeloni et al., 2014).
Having a large number of variable stars covering a wide range of types allows us to use them as references to identify new variables by comparing them to old ones. Angeloni et al. (2014) introduces the VVV template project, which is a database of high quality near-IR light curves for a variety of variable stars. This is still an ongoing project and the numbers of variables used are still significantly smaller than similar databases for visible light curves (Angeloni et al., 2014; Heinze et al., 2018).

One possible way to expand our effective light curve database will be to combine VVV’s near-IR data with optical data from other surveys, which will allow the use of visual light curves for variables from sources such as OGLE-III (Szymański et al., 2011) and ATLAS (Heinze et al., 2018).

Figure 3.1 is an example of how spectrum information - here in the form of multi-wavelength data, can strongly distinguish between objects of similar types.

### 3.2.3 Selection of LPVs

The main objects I am interested in are Long Period Variables (LPVs). Although some of them are famous for being quite obvious, such as the eponymous Mira’s, in the near-IR, it can be harder to select than their fainter cousins. As a continuation of the previous section, this section will describe how Mira’s, Semi-Regular
variables (SRs) and pulsating red giant variables were targeting specifically. To add to this, specific distributions of properties and features of these stars. We will also discuss specific distributions of properties and features of these stars.

**Stellar Properties**

When dealing with evolved low mass stars, there may be dust from the star interfering with the line of sight, adding additional extinction to the existing interstellar extinction. This additional dust is a unknown property and difficult to account for without spectra or access to mid-IR wavebands to probe for IR excess. With the data available, it can only be inferred if other sources of extinction are properly adjusted for.

As discussed previously, the extinction maps of Schultheis et al. (2014) are used. To provide distance estimation, I take advantage of the VVV’s existing crossmatch with Gaia Early Data Release 3 and it’s available parallaxes. Care must be taken with these parallaxes - as Bailer-Jones et al. (2021) note, they have a non-linear relationship with distance in addition to noise that increases with distance.

Some of the objects with Gaia parallaxes are clearly not accurate - for example negative parallaxes. These can arise, typically in noisy observations, when an object does not fit well to the helix-like model for proper motion as the object moves across the sky, where the noise dominates (Luri et al., 2018).

Dealing with this is not trivial. Negative parallaxes lead to nonphysical distances, but removing them biases the sample (Luri et al., 2018). The parallax of an object with small parallax will be more affected by a noisy observation - and more chance that the overall parallax will be brought into the negative by said noise. This is less likely to occur for an object with high parallax. Simply removing all objects with negative parallax will remove more small parallax objects - the sample favouring closer stars with higher parallax. Similar cuts of removing objects with negative parallax in addition to objects with parallax errors beyond a given threshold have the same effect; favouring the removal of stars with small parallaxes. Truncation via limits in parallax, even for thresholds above zero, only lessens the bias.

Fabricius et al. (2021) recommend that when selecting a sample of positive parallax objects, supplement that with an similar sample of objects but with negative parallaxes. With both, test for the fraction of spurious results and
use that knowledge to include the valid objects with negative parallaxes. They also recommend using the re-normalised unit weight error for astrometry data (RUWE), high values of which denote objects that have not been properly resolved (which can also include a lot of binaries).

The recommended solution to this (Luri et al., 2018) is to treat the derivation of parameters from astrophysical data as an inference problem. This is done by Bailero-Jones et al. (2021), who provide a catalogue of stars with distances more accurate than by using parallax alone. By combining the parallax, its uncertainty and a prior, it is possible to calculate a posterior probability distribution (Bailer-Jones, 2015).

The 2021 paper, in addition to using the newer and accurate parallaxes, uses the colour and magnitude of the star in addition to the parallax in order to improve the relationship with the prior. The data to construct this prior comes from the Gaia Early DR3 mock catalogue (Rybizki et al., 2020), containing positions, distances and extinctions for over a billion simulated stars. These parameters were used for this model. The authors note that the priors for this are calculated at higher galactic latitudes, meaning they will not necessarily be an accurate fit in the lower latitude regions with higher extinction and more objects along the line-of-sight. In this case, they advise that the distances calculated without using colour and magnitude information should be used.

Examples of the predicted distances for random samples of 1000 OSARGs, SRVs and Miras found by crossmatching OGLE objects with VVV and crossmatching those VVV objects to Gaia DR2 objects with parallaxes are shown in Figure 3.2.

A comparison to Figure 3.2 demonstrating the difference in predicted distances is shown in Figure 3.3. Both distance methods predict high distances to our objects quite uniformly, although as shown in Figure 3.4 the photometric distances estimates the distances of the OSARGs to be further away than the Miras - given that OSARGs are less evolved and fainter stars, observing them more at fainter distances seems erroneous. The purely geometric distances have more in common across each object type but the overall trend of OSARGs being observed more frequently at lower distances, since they are fainter and will not be detected at the distances many Miras can, is somewhat what was expected. In both methods, the distributions for all objects peak in the 7 kpc to 8 kpc range. All the objects listed here are bulge objects provided by OGLE (Wyrzykowski et al., 2015); if said objects lie at 7 kpc towards the galactic centre, they are probably quite
Figure 3.2  Samples of 1000 Miras, SRVs and OSARGs (Top, Middle, Bottom) plotted according to their VVV apparent magnitude against their Bailer-Jones et al. (2021) distances using photometric information. The error for each point, both in terms of magnitude and the upper and lower distances predicted are shown.
Figure 3.3  Samples of 2000 Miras, SRVs and OSARGs (Top, Middle, Bottom) plotted according to their VVV apparent magnitude against their Bailer-Jones et al. (2021) distances, this time calculated only using geometric information. The error for each point, both in terms of magnitude and the upper and lower distances predicted are shown.
Figure 3.4  Photometric (Top) vs Non-photometric distances (Bottom) (Bailer-Jones et al., 2021) compared for samples of 2000 Miras, SRVs and OSARGs. The height of any point describes the fraction of each object contained at that distance.
affected heavily by extinction.

In a sample of 5000 OSARGs the average parallax uncertainly was 0.3 milli-arcseconds (mas), corresponding to an average photometric distance estimation variance of \( \sim 1900 \) parsecs further and \( \sim 1600 \) parsecs closer. If the geometric distances are used instead, amplitude of variation increases by \( \sim 20\% \). Although using the photometric distances does not prevent the issue from occurring, I believe use of the photometric over geometric distances is an obvious choice.

It is uncertain if Gaia would be able to resolve a object towards the galactic centre behind 7 kpc of dust. If the object lay at a lesser distance, it would make sense that the photometric distance would predict a greater distance because of extinction reduced magnitude, but the purely geometric distance does not take into account the magnitude.

Using these distances and extinction, the training sample of variable stars - found using crossmatching with OGLE - are used to correct for extinction and examine where the variable sample lies on the Hertzsprung-Russell diagram.

The colour vs Ks magnitude of galactic Miras was compared with the work of Matsunaga et al. \( (2017) \) and van Loon et al. \( (2003) \) in Figures 3.5 and 3.6. There is a very noticeable difference between the cluster with the literature Miras being far brighter with the exception of the Mira found using 2MASS photometry, which agree in magnitude. Both disagree in colour, with the VVV Mira being bluer. Figure 3.6 does make use of a 4 magnitude alteration \( (\text{Matsunaga et al.}, 2017) \) in order to compare their Miras to the OGLE LMC ones, I do not believe any issues with this correction that would explain the colour difference or the magnitude discrepancy.

The same colour vs Ks magnitude was repeated again, but this time with absolute magnitudes in Figure 3.7. The absolute magnitude, defined as the apparent magnitude an object would have if it was 10 pc away, can give strong insight into the evolutionary state of an object as it can be used as a proxy for an objects luminosity. It is calculated using the following equation:

\[
M = m - 5(\log_{10} d - 1) \quad (3.1)
\]

where \( M \) is the absolute magnitude in a given band, \( m \) is the apparent magnitude.
Figure 3.5  Colour (J-Ks) vs. Ks Magnitude (Corrected Mag, adjusted for extinction) diagram for OGLE-sourced Miras in VVV, coloured by number of objects. Extinction for both J and Ks is corrected using the extinction maps of Schultheis et al. (2014) using the objects position on the sky and distance from the Bailer-Jones et al. (2021) parallax distances.
Figure 3.6  Colour-Ks Magnitude diagram for Miras. The circles and empty squares denote C-rich and O-rich Miras respectively. The pluses and stars are objects using 2MASS photometry. The grey dots and crosses denote O-rich and C-rich OGLE LMC Miras respectively, which have a correction of 4 magnitudes to simulate their potential bulge magnitudes. Figure 6 from [Matsunaga et al. (2017)]
in a single band and d is the distance in parsecs.

Literature values of the absolute magnitudes of Miras have strong dependence on wavelength. Early papers observed the typical absolute K band magnitude of Miras to be between -7 to -8.8, with one outlying Mira at -5.5 (Robertson & Feast, 1981). In Figure 3.9 Sudou et al. (2019) plot the absolute K band magnitude versus period of a sample of LPVs and Miras, including their predictions of where these LPVs can exist in this parameter space by converting the period-magnitude sequences of Ita et al. (2004) to absolute magnitude. Their equations predict -9 as the peak K band absolute magnitude with even the bright Mira falling below absolute magnitudes of -7 at periods of ~100 days. Absolute magnitudes of these stars in the visual has been calculated measured between 1.0 and -5.4 mag in the Gaia G band (Gigoyan, 2020).

The absolute magnitude of SRVs is similar. Sudou et al. (2019) found that the K band absolute magnitude of the SRV star SV Pegasus was $M_K = -8.09 \pm 0.05 \text{mag}$, whilst Nakagawa et al. (2018) observed absolute magnitudes of $\sim -8$. Nakagawa
Figure 3.8  The absolute $K$ magnitude against log period for a sample of Miras in dwarf spheroidals in NGC 6822. The points are colour coded as Fornax (cyan), Leo I (magenta), Sculptor (green) and Phoenix (black). The black trend line is the PL relation for the LMC. Figure 6 from Whitelock (2012).
Figure 3.9  The absolute $K$ magnitude against log period for a sample of local Miras (squares) and SRVs (triangles). The labelled diagonal regions (B+, C’ and C) denote reproductions of the period-magnitude sequences of Ita et al. (2004). B+ describes the upper part of the sequence for SRV stars and C’ and C denote Mira variables. Figure 9 from Sudou et al. (2019).
et al. (2018) also studied four C-type SRVs - denoting red supergiant stars, far brighter (absolute magnitudes $\sim -11$) and with multi-year periods. This last group is likely rare, given their supergiant nature and mass.

A similar effect to the apparent magnitude is noticed here with the absolute magnitude - a sharp 2 Ks mag decrease in brightness between the literature Mira and the Mira in VVV. One explanation is that the errors on many of the Bailer-Jones distances for these LPVs are rather high. This could potentially be caused by the high mass-loss and resulting dust (Perrin et al., 2020) and large size of these stars distorting the parallax and adding additional uncertainty where there was none. However, a magnitude discrepancy of 2 Ks mag would have to come from underestimating the distance by over half - which seems unlikely if, as noted earlier, the distances predicted for these objects are too great already.

The above examination was repeated for both SRVs and OSARGs and can be seen in Figures 3.10 and 3.11 for Semi-regular variables and in Figures 3.12 and 3.13. The same phenomenon occurs, for reasons unknown.

The cutoff point for RGB stars vs AGB stars is the tip of the red giant branch (Ita et al., 2002). Ita et al. (2004, 2002) studied variable stars in the LMC, using OGLE’s classification to select the variable stars. They noticed that in the K band magnitude there was a noticeable cutoff at 12.2 mags which was identified as the tip of the red giant branch (TTRGB). Above this were the intermediate age AGB stars and early AGB and other RGB stars could be found just below it. By making use of the known distances to the LMC - 49.59 ± 0.54 kpc (Pietrzyński et al., 2019) we can predict that the TTRPG occurs at an absolute K band magnitude of roughly -6.2.

Sudou et al. (2019) compared the distance predicted by the Gaia DR2 parallaxes for the SRV star SV Pegasus to their own calculated parallaxes, measured using the positional changes of $H_2O$ maser spots plus near-IR monitoring. There is strong difference in parallax measured; 3.00 ± 0.06 milliarcseconds for Sudou et al. (2019) versus 1.12 ± 0.28 milliarcseconds for the Gaia measurement. The authors suggest that the large size of the SRV is the cause - an estimated radius of 5 milliarcseconds for the 333 pc away object. This circumstances for this case are quite different to ours - we certainly won’t be observing a LPV at these close distances with mean K magnitudes of $\sim -0.5$ mag, but the variation of the object at close distances prevents accurate parallaxes and the further away objects suffer from increasing parallax errors with increasing distance (Xu et al. 2019),
Figure 3.10 Colour (J-Ks) vs. Ks Magnitude (Corrected Mag, adjusted for extinction) diagram for OGLE-sourced SRVs in VVV, coloured by number of objects. Extinction correction is performed as described in Figure 3.5.
Figure 3.11  Colour (J-Ks) vs. Ks Absolute Magnitude diagram for OGLE-sourced SRVs in VVV, coloured by number of objects. Extinction correction is performed as described in Figure 3.5.
Figure 3.12  Colour (J-Ks) vs. Ks Magnitude (Corrected Mag, adjusted for extinction) diagram for OGLE-sourced OSARGs in VVV, coloured by number of objects. Extinction correction is performed as described in Figure 3.5.
Figure 3.13  Colour (J-Ks) vs. Ks Absolute Magnitude diagram for OGLE-sourced OSARGs in VVV, coloured by number of objects. Extinction correction is performed as described in Figure 3.5.
hindering accurate parallaxes in both ranges. Xu et al. (2019) examine high-error parallax AGB stars, citing the objects large size, surface brightness variations and circumstellar dust shells as the cause. The authors suggest caution when using said Gaia parallaxes, rather than discounting them completely. In a similar vein, Sun et al. (2022) observed a $3.8\sigma$ tension between their maser-determined parallaxes and Gaia EDR3 parallaxes and also concluded that the photocenter variations play a large part.

Figure 3.14 demonstrates the positions of the candidate variables I am working with. The distances predicted for these LPVs are not impossible but it seems unlikely that there are such numerous groups of objects at relatively high inclinations, whilst also observing closer objects much more infrequently. The above distribution also occurred for both Mira and OSARG LPVs. It is a strong possibility that distance band is strongly influenced by the prior used by Bailer-Jones et al. (2021), as the authors note that for poor data or high parallax uncertainty the posterior will be dominated by this prior. A confusing aspect of this explanation is why the prior for this galactic bulge region would be $\sim 7$ kpc - a combination of distance and extinction that Gaia would struggle to observe through.

Even though the distances are faulty, they cannot alone explain the discrepancy in absolute K band magnitude observed. Many of them appear to have too low absolute magnitudes at their average distances of $\sim 7$ kpc, but distances that would explain a discrepancy of multiple magnitudes would require these objects to be far more distant than this - several tens of kiloparsecs or more. It is not out of the question that a RGB or AGB star could be observed at such distances, but if this were true would OGLE have observed them in the first place?

Unaccounted for extinction cannot be the sole cause either. Using the Schultheis et al. (2014) extinction map described in Section 2.1.2 for the same sample of objects as before, the mean Ks band extinction is $\sim 0.3$, which although non-zero does not account for the several magnitude discrepancy. If the variable stars collected do have extinction any higher than this they will be experiencing optical extinction in the realm of several magnitudes - enough to hinder finding counterparts in OGLE and Gaia. Since finding said counterparts in OGLE and Gaia was successful, it is unlikely that this is the case. It is also unlikely that both factors are at play, for the same reasons. An object both distance and with moderate extinction will struggle to have matching counterparts in the two optical surveys I am comparing it with.
Figure 3.14 From a sample of 1000 SRVs originally detected using OGLE, with counterparts in VVV and Gaia EDR3 distances using Bailer-Jones et al. (2021). Plot 1 denotes objects plotted according to their galactic latitude by their distance. Plot 2 denotes objects plotted according to their altitude above the galactic plane with respect to their distance. The colour of each point denotes the galactic longitude, relegated to a third axis since I wanted to better understand what inclinations the ∼ 7 kpc predicted objects (Bailer-Jones et al., 2021) lay at.
The possibility of incorrect matching is one of the few other explanations I considered. When matching between the OGLE LPVs and the VVV objects in moderately dense regions, confusion between objects could occur. However, I predict that when there is confusion and overlap between multiple sources along a sight line, the brightest object should dominate. The brightest object would then be the counterpart considered for each match. Since we are working with LPVs which are luminous RGB and AGB stars and bright in the near-IR, the LPVs will be brighter than non-evolved stars at the same distance and closer stars \( \frac{1}{10} \) of the distance. As such, I do not believe that close light-of-sight objects provide a reasonable explanation for the discrepancy either.

At this stage, I am uncertain what the cause of this phenomenon is.

**Binarity**

Binary stars are incredibly common - as so it should not be a surprise that a star could be both in a binary (or higher) pair and also show signs on pulsating variability (Karczmarek et al., 2022). Current estimations put the percentage of stars with binary companions at: low-mass stars at \( \sim 50\% \) (Raghavan et al., 2010) and high-mass stars at \( \sim 70\% \) (Sana et al., 2012), so it should be expected that some of the LPVs in our sample will also display multiple variations. Current works have already begun studying Cepheids with binary companions - currently a third of galactic cepheids are known to have companions (Neilson et al., 2015). RR Lyrae have been observed in binaries, but the sample with observed partners is much smaller (Hajdu, 2021), with only a handful of cases, believed due to their lower mass (sub-solar) (Kervella et al., 2019). From a physical standpoint, short period binarity is more likely to be observed (Hajdu, 2021) - it is more likely that an observation catches the multiple stars overlapping if the stars orbit more frequently.

The scale of the variations is an additional consideration. The majority of the dataset used to train are presumed to, if they have binary components (aside from objects listed as binary), be the larger and brighter components of their systems. As LPVs, these stars are evolved, big and bright, so any brighter binary companion would be the main star and likely dominate the pair. Consequently, the effect of a small companion will be also be small, with the typical variations on the scale of a few percent in brightness (Miszuda et al., 2021). This should not be too big of an issue for this project, given all of the objects we are most
interested in are LPVs, giant stars with larger amplitude than other pulsating variables.

### 3.2.4 Variability Indices

Various methods of selecting candidate variables from a sample have been developed over time, with many methods calculating statistics that characterize the time domain light curve, and then using these measures in detection and classification. Viable statistics include the maximum, minimum, mean, variance, kurtosis and median absolute deviation, but it is also common to use computed test statistics and variability indices. These latter include the chi-squared test, the Welch-Stetson variability indices \cite{Welch & Stetson 1993} or the Con measurement \cite{Wozniak 2000}.

Variability indices are properties that can be used to "check" if an object is variable, in addition to what type of variable. These are mathematical values which vary uniquely between objects - some objects will have very similar periods, other objects will have very similar temperature, but by using both, a better distinction between types can be drawn. I will describe the properties like periods, light curve statistical properties like skew and information from and about the environment and how they can best be used to extract maximum information.

This section will include the ones I have used, some of their derivations if needed, and the Principal Component Analysis (PCA) used to select only the ones that will give the best performance.

The choice of input features for the classifier should depend on what is available and the goals of the classification. Some techniques are better suited to detect certain kinds of variability than others, such as using metallicity to distinguish between Delta Scuti and SX Phoenicis variables \cite{Kopacki 2007} and properties based on light curve shape to identify Eclipsing Binaries. Multiple statistics are combined to obtain a better return of variable objects \cite{Heinze et al. 2018, Kim et al. 2011} which leads to an overall superior detection rate \cite{Sokolovsky et al. 2017}. For this reason in this project periodograms and variability indices are both used to select variables, using the approach of \cite{Ferreira Lopes et al. 2020}. This approach is well suited to an approach involving machine learning, which can decide on decision boundaries for multiple statistics and adjust them as needed.
Some approaches use a period fitting analysis as the variable selection process. If an object has a period fit above a certain threshold of goodness-of-fit, the object is flagged as variable. See section 2.3.2 for a discussion on period finding algorithms. This method is well tested, being used to select variables in multiple surveys with good success (Jayasinghe et al., 2018; Heinze et al., 2018). The use of periodograms as a detection technique is more computationally expensive than time series analysis methods like the Welch-Stetson indices, but is more sensitive to low amplitude variables (Heinze et al., 2018). However as computing power increases, this method becomes more viable as one of the (or the initial) selection process. Additionally, periodogram methods can detect variability of non-periodic variables, as long as the variability has time-coherent behavior (Heinze et al., 2018).

There have been comprehensive reviews of various indices that could be used to select objects believed to be variable. Shin et al. (2009) consider 6 variability indices. Sokolovsky et al. (2017) include more in their review and make several distinctions in indices, separating them into methods based on the scatter between magnitude points without considering time information present along with those points, and correlation based methods which consider the order or time of the measurements. Pashchenko et al. (2018) considered 24 different indices while noting that some of these are measures of the same quantity of the data, but with different weighting or other small statistical effects. They used the Pearson Correlation Coefficient, a measure or linear correlation between two datasets (F.R.S., 1901), to remove those indices that were too correlated with other existing indices, favoring the less computationally expensive and most informative indices.

Dubath et al. (2011) performed a random forest classification of Hipparcos variable stars and determined the attributes that encapsulate as much data as possible, whilst removing attributes that are heavily correlated with other attributes but overall provide less information to narrow down to a list of most important attributes. This list includes period, amplitude, colour, absolute magnitude and skewness. The period was calculated as the most important aspect, and any objects with incorrect periods calculated were found to be misclassified approximately twice as likely as objects with correct periods (Dubath et al., 2011). This list of attributes has been used to classify variables by the Gaia Collaboration (Eyer et al., 2017). Kim & Bailer-Jones (2016) likewise find that period is one of the variability indices with the most impact on a successful or unsuccessful classification.
The features I considered (alongside Absolute Magnitude and Colour, both simple enough that they need no introduction) were are described below.

**Period and Temperature**

The period and temperature are described in Sections 2.3.2 and 2.3.1.

**Amplitude**

Defined as half the difference between the maximum and minimum magnitudes. To improve how reliable this statistic is, outlying points must be handled carefully. An erroneous outlying point with high or low magnitude will swing the result artificially. This is often solved by using the 95% and 5% percentiles instead of the max or min. This serves to trim outlying points whilst still keeping the same information.

\[
A = \left( \frac{m_{\text{max}} - m_{\text{min}}}{2} \right)
\]

where \(A\) is the amplitude and \(m\) in the magnitude of the object. The max and min magnitudes of a VVV object are part of the Vista Science Archives (VSA) vvvVariability table (Cross et al., 2009) as vvvVariability.ksMaxMag and vvvVariability.ksMinMag respectively for Ks-band magnitudes.

**Skew**

An index of the asymmetry of a distribution (Friedrich et al., 1997), calculated using the third order central moment and the variance (D’Isanto et al., 2016).

\[
Skew = \frac{\eta_3}{\sigma^3}
\]

where \(\eta_3\) is the third central moment and \(\sigma^3\) is the variance. The skew is pre-calculated in the VSA archives in vvvVariability table as vvvVariability.ksskewness.
Kurtosis

The Kurtosis describes the departure of a distribution from the norm (Friedrich et al., 1997). The equation given below holds describes the small kurtosis, an approximation that works best when the number of epochs is low (D’Isanto et al., 2016).

\[
\text{Kurtosis} = \frac{\eta_4}{\sigma^2}
\]  

(3.4)

I calculated the kurtosis from the VVV light curves using the 2” radius aperture Ks magnitudes for each detection (vvvDetection.aperMag3).

Standard Deviation Ratios

The ratio of standard deviation to the sample mean magnitude

\[
\frac{\sigma}{\mu} = \sqrt{\frac{\sum_{n=1}^{N} (x_n - \mu)^2 / (N - 1)}{\sum_{n=1}^{N} (x_n / N)}}
\]  

(3.5)

where \( x_n \) is the \( n \)th point in a light curve out of \( N \). A high value indicates likely variability (Shin et al. 2009). I calculated these ratios using vvvDetection.aperMag3 magnitudes.

Con

A feature tallying the number of 3 or more consecutive points at least 2 \( \sigma \) above the mean magnitude (defined above). This tally is normalized by dividing by \( (N-2) \) (Wozniak 2000). I calculated these using vvvDetection.aperMag3 magnitudes.

von Neumann ratio

The ratio of mean square to the successive difference (difference between a point and its neighbour) to the variance (von Neumann 1941).
\[ \eta = \frac{\delta^2}{\sigma^2} = \frac{\sum_{n=1}^{N-1} (x_{n+1} - x_n)^2}{\sigma^2(N - 1)} \] \hspace{1cm} (3.6)

often known as the von Neumann ratio. Strong correlation or anti-correlation lead to low or high values respectively \cite{Shin2009}. I calculated this ratio using vvvDetection.aperMag3 magnitudes.

**Chi-squared**

\[ \chi^2 = \sum_{i=1}^{N} \frac{(m_i - \tilde{m})^2}{\sigma_i^2} \] \hspace{1cm} (3.7)

where \( \tilde{m} \) is the weighted mean magnitude \cite{Sokolovsky2017} and \( m_i \) is the \( i \)th magnitude in a series of length N.

\[ \tilde{m} = \frac{\sum_{i=1}^{N} m_i}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}} \] \hspace{1cm} (3.8)

and the resulting \( \chi^2 \) is compared to a \( \chi^2 \)-distribution corresponding to the null hypothesis. The Chi-squared is pre-calculated in the VSA archives in vvvVariability table as vvvVariability.chi2.

**Mean Absolute Deviation (MAD)**

\[ MAD = median(|m_i - median(m_i)|) \] \hspace{1cm} (3.9)

is used as a measure of scatter \cite{Pashchenko2018}. The MAD is pre-calculated in the VSA archives in vvvVariability table as vvvVariability.ksMagMAD.

**Interquartile range (IQR)**

The interquartile range is defined as the inner 50\% of a dataset, excluding the
highest and lowest 25%. I calculated it using vvvDetection.aperMag3 magnitudes. The technique has notable usage in the study of objects with asymmetric distribution, such as the light curves of binary systems [Sokolovsky et al., 2017].

**Robust Median Statistic (RoMS)**

\[
RoMS = \frac{1}{N-1} \sum_{i=1}^{N} \frac{|m_i - \text{median}(m_i)|}{\sigma_i} 
\]

(3.10)

has a value of 1 for objects where the majority of the values fall within 1 \(\sigma\) of the median. I calculated this using vvvDetection.aperMag3 magnitudes.

**Stetson Indices**

These indices and the following were developed by the astronomy community, with Stetson J and K developed by Stetson (1996).

\[
J = \sum_{n=1}^{N-1} \text{sign}(\delta_n \delta_{n+1}) \sqrt{\delta_n \delta_{n+1}} 
\]

(3.11)

\[
K = \frac{1/N \sum_{n=1}^{N} |\delta_n|}{\sqrt{1/N \sum_{n=1}^{N} \delta_n^2}} 
\]

(3.12)

where

\[
\delta_n = \sqrt{N/(N-1)}(x_n - \mu)/e_n 
\]

(3.13)

where \(e_n\) is the photometric error for each point.

Welch-Stetson variability index measures the degree of correlation between pairs of measurements [Welch & Stetson, 1993].
\[ I = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{N} \frac{b_i - \bar{b} v_i - \tilde{v}}{\sigma_{b_i} \sigma_{v_i}}} \]  

(3.14)

where \( b_i, \sigma_{b_i} \) and \( \tilde{b} \) are the measured magnitudes, errors on said magnitudes and the mean of said points and the corresponding values for \( v \) are for another point close in time to the first. The measured magnitudes continue to be vvvDetection.aperMag3 magnitudes as before.

This indices has had variants created over the years, such as Fruth et al. (2012)'s time weighted version, which treats points differently by their closeness in time. The author noted their technique as being particularly superior to the original in cases where the amplitude of variation is comparable to the noise level. Since this is not the case for my objects of interest, I defaulted to the simpler to calculate original indices.

**ANOVA**

The last indices used by Shin et al. (2009) is the Analysis of Variance (ANOVA) statistic (Schwarzenberg-Czerny, 1996). The derivation of this statistic is not included due to its length. It is used by separating an object into the different components that make up its variation, and typically comparing the means between the components. I calculated this statistic using built in methods as part of ScyPy (Virtanen et al., 2020).

An overview of what indices have been considered useful by modern variability reviews and studies is shown in Table 3.1.

**Viva Indices**

With the use of the VIVA catalogue there are many additional indices available to use. The parameters calculated by Ferreira Lopes & Cross (2016), Ferreira Lopes & Cross (2017) and Ferreira Lopes et al. (2020) to select the variables this project plans to use are a variant of the Stetson indices (Stetson, 1996). These indices are shown to have A good F-score when tested on multiple variable star catalogues (Sokolovsky et al., 2017), but it is noted that the indices performs less
Table 3.1  A tallying of the variability indices used in modern astrophysics studies and reviews. This is not a complete sample of every index used and some have been combined with others for the sake of brevity. The standard deviation $\sigma$ is used by itself as well in more complex indices such as $\frac{\sigma}{\mu}$. The von Neumann ratio is often used as $\eta$. The Stetson indices have clipped and time weighted versions described in their section that are merged with the base version. The references are aliased as follows: 2: Shin et al. (2009) 3: Sokolovsky et al. (2017) 4: Pashchenko et al. (2018) 5: Sokolovsky & Lebedev (2018) 6: D’Isanto et al. (2016)
well on objects with the occasional change in brightness interrupting a normally constant brightness. This includes flare stars and EA binaries stars.

Correlation-based indices like the Stetson Indices improve in performance as the number of epochs available increases, whereas scatter based measurements maintain constant performance even at low epochs (Number of epochs \( \approx 50 \)). The trade off is higher performance when more epochs are available, but some indices including Stetson, still show reasonable performance on under-sampled data (Number of epochs \( \approx 100 \)) \cite{Sokolovsky2017}.

In the original paper, \textcite{FerreiraLopes2015} modified the Welch-Stetson indices to quantify panchromatic flux correlations (pfc) as the index \( I_{\text{pfc}}^{s} \), instead of pairwise flux correlations \( I_{\text{WS}} \). They also proposed a flux independent (fi) index \( I_{f}^{s} \) which is less insensitive to true variables with one to several outlying data points and poor error estimations.

In a later paper, \textcite{FerreiraLopesCross2016} proposed even more robust and powerful indices. Where \( I_{\text{pfc}}^{s} \) would not perform as well for different numbers of observations in each group observed over a short time period, the index \( K_{f}^{s} \) which describes the ratio of positive correlated groups to the total and can work on single filter data unlike \( I_{f}^{s} \), the index \( L_{\text{pfc}}^{s} \) which now works on single epoch data and reduces to \( J_{\text{pfc}}^{s} \) in the one filter case. The remaining indices are the median number of correlations \( M_{\text{pfc}}^{s} \), a correction factor related to outliers and instrumental factors \( F^{s} \) and two combined indices \( FL^{s} \) and \( FM^{s} \). For these indices, \( s \) denotes the number of wavebands being compared for correlation in each calculation step, which can take a maximum value equal to the total number of wavebands present in the data.

As mentioned above, indices based on correlation require time information. For the relevant indices, this means that the "boxes" in which the data points within can be considered the same epoch must have their size in time set. Too large and variability begins returning values of the indices closer to what would be expected for noise. These indices were tested on how effectively they found 80% and 90% of the variable stars in the Wide Field Camera Calibration (WFCAMCAL) Variable Star Catalogue (WVSC1) \cite{FerreiraLopes2015}. The most effective of these \( K_{f}^{(3)} \) returned three times as many sources in the selection compared to the known variables in the WVSC1 in order to enclose 90% of the WVSC1 stars, for all stars with more than 20 epochs of data available. Whilst some further selection is needed, this method shows excellent accuracy in selecting variable objects and
will be the initial selection tool used in this project. The final decision of what indices are used is described in more detail in Section 3.3.

The machine learning algorithms are capable of finding patterns in these indices and combinations of indices, but a preliminary investigation can also be done manually. Doing so can help to determine which indices contain the most information, which contain little, and generally improve our understanding of the data available.

**Principle Component Analysis (PCA) and other indices distribution analysis**

PCA reduces down the dimensionality of the data, encapsulating much of the meaning of multiple different indices to a lesser number of parameters with fewer dimensions but with as much information preserved as possible. The section of the VIVA catalogue with known classifications was crossmatched with Gaia to include some optical information, along with using the Gaia magnitudes to obtain an estimate of temperature for each object, as described in section 2.3.1.

In Python, Scikit-Learn (Pedregosa et al., 2012) was used to create a 3 component PCA transform using the following indices as the base parameters to be transformed: Z, Y, H, J, Ks mean magnitudes (VVV), G, Gbp, Grp mean magnitudes (Gaia), mean Ks errors, Ks band amplitude and effective temperature. In addition to these, the following values from the existing VIVA catalogue were used: Periods calculated using the 5 methods included in VIVA: LSG, PDM, STR, Lfl2, PKfl2, even dispersion parameters ED, kFi2, L2 and X variability index Xindex.

The resulting impact of each input on the three principle components (output from the PCA) are shown in Figure 3.15. It portrays how correlated or-anti correlated the base parameters were when creating the 3 PCA components. The period are all correlated together as a group, indicating that they mostly agree and are a strong method for maximum separation of the data points. The temperature is anti-correlated with the periods - indicating that a star with a high period will have a low temperature. Strangely, the periods and extinction also correlate, which could be due to the OGLE sample of Mira and SRV. These long period stars would originate from the OGLE bulge survey which would explain the high extinction for these objects.
Figure 3.15  The impact of each indices from the crossmatched VIVA-Gaia dataset on distinguishing between the various categories. The colour scale denotes the amount of positive correlation (yellow colours) and anti-correlation (blue colours).

To provide a comparison of dimensional-reduced data with respect to the different types of object in the sample, t-distributed stochastic neighbor embedding (t-SNE) \cite{van2008visualizing} was implemented using Scikit-Learn again. This technique preserves the local structure of the data and is useful for visually comparing the distributions of our data classes. The technique reduces and ‘compacts’ the data, modelling it by points closer to a given point being more similar to it (with regards to its higher dimensional properties) and points far away being more dissimilar. This technique was initially designed for unsupervised classification and clustering data points and was not designed with astronomical data in mind, but has nonetheless been put to use finding spectra from binaries and other atypical sources amidst large (300,000) collections of spectra from GALAH \cite{Traven2017} and visualizing the results from predicting AGN types from spectra \cite{Peruzzi2021}.

An example of a two component t-SNE clustering using the same information as the above PCA is shown in Figure 3.16. From it, several things are apparent; Whilst there is a defined region in which the LPVs exist in, it is not abundantly clear what the defining characteristics of such a region are. The eclipsing variables, on the other hand, cover the entire parameter space and whilst some structure is visible, there is clear overlap with the other types of variable. This is to some extent expected - eclipsing variables can occur at low, medium and high
masses and with widely different orbital dynamics.

If structure observed is potentially spurious, a reliable way of confirming it is to test if said structure occurs at other perplexities - one of the main hyperparameters for the algorithm. Perplexity is a tunable parameter with rough correspondence to the size of the subgroups and the density of the dataset. Figure 3.17 demonstrates another example of a plotted t-SNE with different hyperparameters. Whilst some difference is visible, the structure is not clear.

The following representation potentially could have been improved to show clear structure through extensive tweaking of hyperparameters. The machine learning algorithms employed will, by default, fit to the data and be much clearer about how accurate the model is at fitting the data. They also are far more replicable - the machine learning model once trained can be applied to other data, whilst the clusters created above are difficult to apply to other datasets.

3.3 Probabilistic Random Forest (PRF)

Random forest are a category of models with extensive use in astronomical surveys. Sesar et al. (2017) used a gradient tree boosting technique (Friedman 2001) to identify RR Lyrae variables from the Pan-STARRS1 3π Survey; Hernitschek et al. (2016) use a more standard random forest classifier. Yong et al. (2018) compare a random forest with decision trees and logistic regression
to analyse quasars and found that the random forest performed better than the other methods.

The algorithm, created by Breiman (2001b) is simple to use, and as mentioned above, shows high performance (Reis et al., 2019) and is used for both classification - applying labels to unknown objects and regression - involving estimation or prediction of properties of an objects.

**Algorithms**

The algorithm is an ensemble learning method, taking the results of many separate threads (in this case, decision trees) and using the aggregate decision of each tree to make a decision (Breiman 2001b).

Each decision tree describes its relationship between its input and output as a series of splitting paths, starting with an initial split and the branches after doing the same, resulting in a tree-like graph. Each of these conditions used to create a split is of the form $X_j > x_{j,th}$. This equation denotes how, for a input feature of the tree j, if the value of said input feature exceeds a threshold value $x_{j,th}$, the input proceeds down one branch and if the input feature does not exceed the threshold, proceed down the other branch (Reis et al., 2019).

The feature used as well as the threshold value are both set during training, during which the algorithm tries to identify what the combination of feature and
threshold leads to the greatest separation (Breiman, 2001b). The definition of greatest is a hyperparameter of the model. One example measure for determining greatest separation is the "Gini impurity" - defined for a group as the probability of a randomly selected object being misclassified (Reis et al., 2019).

A single tree is prone to overfitting (Breiman, 2001b) - good performance on the data it is trained on and much poorer performance on newer, unseen data. By combining multiple trees, each trained on different subsets of the training data and using different combinations of features for each, the forest avoids overfitting and performs better overall (Breiman, 2001b).

Tree based models are robust to features (measurable properties of the observed object, such as colour or period) that may not be directly useful as classification features. Where other models would be hindered if an addition only-sometimes useful parameter is added, when building the classification model of a tree based model, this can be done without hindering it’s performance, and can still potentially lead to improved classification via correlations that by themselves are not individually more useful than other features (Sesar et al., 2017).

Reis et al. (2019) suggest an improvement. In their modification to the random forest, named the Probabilistic Random Forest (PRF), they design it to make use of the fact that there will be uncertainties in all measurements we take and properties we calculate. The normal random forest only uses the values themselves - taking an input of

\[(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\]

where x is the set of input data (multiple features) and y is the assigned label. These obey some unknown relation \(h(x) \rightarrow y\), which the random forest attempts to model, which can then be used to predict the label y for any new set of x. The PRF receives different input

\[(x_1, y_1, \Delta x_1, \Delta y_1), (x_2, y_2, \Delta x_2, \Delta y_2), \ldots (x_n, y_n, \Delta x_n, \Delta y_n)\]

where the \(\Delta x\) and \(\Delta y\) are the uncertainty in the input features and the uncertainty in the labelling respectively (Reis et al., 2019). The PRF learns the relationship \(h(x, \Delta x) \rightarrow y(\Delta y)\), which the network models, using the uncertainties on the labels to help it understand the overlap between the labels it sees. This trained model can then be used with sets of \(x, \Delta x\) to predict the label of y. The idea of input parameters having uncertainty is well understood - when dealing with limited data
Figure 3.18  A graphical example of the difference between the nodes of a random forest tree (upper) and those of a PRF (lower), showing the process for an object propagating through the node. The random forest makes a simple boolean decision with one of two outcomes. The PRF sends the down both paths with a different weighting on each - based on the uncertainty of the feature used for that node. Figure 1 from Reis et al. (2019)

and noise that cannot be completely corrected for, errors in magnitude and false spikes in periodograms. The labels having uncertainty is less intuitive, but the hard boundaries we draw are not always physical. Community labelling projects such as Galaxy Zoo (Lintott et al., 2008) or objects observed by multiple sources can lead to objects potentially belonging to multiple labels.

To utilize the uncertainties, the features and labels are be treated in a probabilistic manner. By using probability distribution functions (PDFs) instead of fixed constants to represent the features and labels, we can denote the most likely value of a feature as the peak of a PDF with spread set by the uncertainties (Reis et al., 2019). An example of this is shown in Figure 3.18.

Figure 3.18 also demonstrates how each node of the tree handles the uncertainties on it’s input. At each node, a given object can propagate to both successive nodes. The progress of an input down the tree is shown in Figure 3.19.

As the PRF involves calculations across all the nodes in the tree, there is an increase is training time for this network compared to a regular random forest.
Figure 3.19  Upper, a simple random forest. Lower, a simple PRF. A comparison of the path down an individual decision tree an object takes for each algorithm. In a normal random forest, an object follows a single trajectory down the orange line of nodes and arrives at the final red node outcome. In a PRF, each node propagates it’s input (as well as the probabilities) to each child node, ending up with a prediction at each end node.
Figure 3.20  A trimmed PRF. If the output to a node would fall below an assigned threshold, we can drop that node, removing it from further calculation. This is used to trim nodes where the probability of it being the right one is low enough to remove it, aiding the speed of the network by removing parts it has to calculate. Black nodes and paths denote trimmed pathways.

Figure 3.20 shows a technique to trim nodes that are highly unlikely to be outputs to increase the speed of the algorithm. With the improvement of using the additional data, the algorithm requires less information to be trained. The general robustness of the algorithm over the original is why it was chosen (Reis et al., 2019).

A full mathematical description of the PRF is not included in this paper, please see the original paper by Reis et al. (2019) for a full derivation.

3.3.1 Implementation

The model implemented uses the Python 3 framework of the PRF provided by Reis et al. (2019). The code was adapted with the help of scikit-learn to aid with the splitting of testing and training data, still using Python 3. The network architecture was trained on a 6 core PC with a NVIDIA GeForce GTX 1080 graphics card.

https://github.com/ireis/PRF
3.3.2 Training

The VIVA catalogue was the primary source of objects, chosen by those that also had crossmatch counterparts with Gaia EDR3 data. The Gaia EDR3 crossmatching enabled me to use the [Bailer-Jones et al. (2021)](Bailer-Jones2021) distances calculated from Gaia parallaxes, which were combined with with the extinction map of [Chen et al. (2013)](Chen2013). This was used to correct for the extinction towards the galactic bulge and applied to the following input features:

- **Absolute Ks Magnitude**
  
  - I used the VVV mean Ks magnitude and the [Bailer-Jones et al. (2021)](Bailer-Jones2021) distances to calculate an absolute magnitude, adjusted for extinction using the Bailer-Jones distances and [Schultheis et al. (2014)](Schultheis2014) extinction map. It is important to use this over apparent magnitude - there is risk of the PRF learning to mimic observational patterns instead of object properties. VVV is more likely to observe Miras further away than OSARGs because the Miras are bigger and brighter and would saturate at some distances an OSARG would be observable at. This is an example of an observational pattern.

- **Absolute J Magnitude**
  
  - I used the VVV mean J magnitude and the [Bailer-Jones et al. (2021)](Bailer-Jones2021) distances to calculate an absolute magnitude in the same fashion as the Absolute Ks magnitude. This is included to give the algorithm access to a direct approximation of colour. The errors on these terms were calculated including the uncertainties from the Bailer-Jones distances. I decided that it would better to include high uncertainty distances than try to estimate distances errors. This leads to asymmetrical distances.

- **Amplitude**
  
  - The amplitude of the corrected VVV Ks magnitude light curve data. Errors in max and min magnitude used to calculate errors. This was calculated as defined in Section 3.2.4

- **Effective Temperature**
As described in Section 2.3.1, the IRFM (Mucciarelli & Bellazzini, 2020) is used with the extinction corrected VVV and Gaia magnitudes to calculate an effective temperature. The errors propagated from errors in colour information.

The following features did not make use of the extinction correction:

- **\( K^2_{Fi} \)**
  - This indices from VIVA is calculated using the number of correlations in the Ks band data. As such, it has no dependence on magnitude. The \( K^2_{Fi} \) is pre-calculated in the VSA archives vvvVivaCatalogue.kFi2. I am uncertain how to calculate errors on this quantity. No errors used.

- **\( l^2 \)**
  - This indices from VIVA is calculated using the number of correlations in the Ks band data. As such, it has no dependence on magnitude. The \( l^2 \) is pre-calculated in the VSA archives vvvVivaCatalogue.l2. I am uncertain how to calculate errors on this quantity. No errors used.

- **\( H_{KpK/paPcorrelation} \)**
  - This value is calculated using the period power spectrum height for the Ks band light curves using the flux independent period finding method from Ferreira Lopes et al. (2018) divided by the false alarm probability of the \( K^2_{Fi} \) index. It is an unusual measurement, but Ferreira Lopes et al. (2020) seemed to have good success using it to separate variable from non-variable objects. This value is calculated from the vvvVivaCatalogue.HeightPKfi2 and vvvVivaCatalogue.faPcorrelation2 values in the VSA archives. This property is calculated using periodogram height. I am uncertain how to quantify errors in periodogram spikes. No errors used.

- **Best Period**
  - As described in Section 2.3.2 I calculated a best fitting period by comparing the predictions of several period finding methods. I am uncertain if the extinction corrections causing an adjustment to magnitudes would have an effect on the calculation of period
for an object. I wasn’t certain how to calculate the errors from a periodogram and considered multiple overlapping distributions for each high periodogram spike with a width given by its false alarm probability as a means of establishing confidence in each spike. In the end, I wasn’t sure if this would work. No errors used.

- **xIndex**
  
  The ratio of a statistical parameter, in this case the standard deviation, by its expected noise value. The standard deviation may be affected by the extinction corrections but I lacked the available noise models to recalculate this parameter. As such, the VIVA values were used. The x index is pre-calculated in the VSA archives vvvVivaCatalogue.xIndex. Uncertain how to calculate errors on this quantity. No errors used.

- **wesenheit**
  
  The wesenheit function is specifically derived as a reddening free index \cite{Ngeow2012}, hence no correction was applied. This indices has been used by the OGLE team in their study of LPVs \cite{Soszynski2009, Soszynski2022} and they seemed to have great success distinguishing LPVs using this, which is why I chose this as one of the features. I calculated this using the vvvSource.ksAperMag3 and vvvSource.jAperMag3 values as follows: 
  
  \[
  W_{JK_s} = K_s - 0.69(J - K_s).
  \]
  
  Errors propagated from errors in J and Ks magnitude information.

The training sample consisted of two parts.

The first part, was a sample of variables drawn only from the crossmatched variables available as part of VIVA. Totalling 173875 objects, it consisted of:

- 131380 Eclipsing Variables
- 36168 Pulsating Variables
- 4231 Long Period Variables
- 2096 Young Stellar Objects

Where each type of variable star had been grouped into their broader categories as described in the background \cite{1}. This is a wide variety of objects, but I
believed that the Long Period variables would be detectable due to their long
periods, and training the network with eclipsing binaries would help the network
to remove the possible contamination from the few detached binaries with longer
periods. Pulsating variables were included because they were available and did
not negatively impact performance. Although this project has a focus on LPVs,
including the remaining pulsating variables could be helpful for other researchers
using the VVV data.

The size of the eclipsing sample dominates the other categories and thus a random
third of the eclipsing objects were used in any given run of the PRF, turning the
final numbers for this sample to:

- 39414 Eclipsing Variables
- 36168 Pulsating Variables
- 4231 Long Period Variables
- 2096 Young Stellar Objects

These objects were shuffled within their types and a training sample of 25,000
objects and a testing sample of 25,000 objects were obtained from the total. These
were used in the part one of the training, in the PRF described above. A total
of 100 trees - 100 estimators were used when training the network. The original
paper used 10 trees (Reis et al., 2019), but slight performance increases were
noticed when using 100, and the speed of the algorithm meant that the network
still trained in under 30 minutes.

The second version of the PRF was created with the aim of examine how the
algorithm was affected by using the more specific classes over the broader classes,
with particular regard for the LPVs, to test how well the network can split the
data into its subtypes - if performance is not adequate, the LPVs found by the
first part will be separated by period manually. This could be necessary if the
network prioritizes other indices over the period, on in cases where the best fit
period is not sufficient and the top few periods must be considered in order to
remove noise-caused false periods or genuine multiple periods that occur in a
single object.

Since the sample is already quite unbalanced, with eclipsing and pulsating
variables making up the majority of the objects, any attempt at simply dividing
the 4231 LPVs into their constituent types would bring their numbers down dangerously low, with the number of Miras in the above sample being \( \sim 150 \). To augment these objects, a sample of OGLE LPVs - which had both VIVA and Gaia counterparts, were used to increase the number of each object in the second test to:

- 12760 Eclipsing Variables
- 10850 Pulsating Variables
- 17884 OSARGs
- 699 Miras
- 2614 SRVs
- 2096 Young Stellar Objects

As the numbers of each object are reduced, the training sample and testing sample were both reduced down to 15,000 objects each.

Noticeably, the sample of OGLE objects with both VIVA and Gaia counterparts did not increase the number of Miras by very much. I believe this may be due to a processing aspect of using the VIVA catalogue. A check recommended by the creator (Ferreira Lopes et al., 2020) is to reject objects with insufficient numbers of good data points - below 30. An object visible to both OGLE and Gaia may be too bright, reducing how many valid data points were available for each object.

### 3.3.3 Validation and Results

In order to calibrate the classifier and measure its performance, 3 measurements are used. These are Completeness and Precision, which are the ratios of detected variables to total variables and the ratio of detected variables to detected objects respectively (Kim et al., 2011). These two are combined in an F-score (Graham et al., 2014) that denotes an overall performance score.

\[
Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{3.15}
\]
\begin{equation}
Recall_i = \frac{TP_i}{TP_i + FN_i}
\end{equation}

\begin{equation}
F1 - score_i = \frac{2 \cdot Precision_i \cdot Recall_i}{Precision_i + Recall_i}
\end{equation}

$TP_i$ is the number of true positives, $FP_i$ is the number of false positives and $FN_i$ is the number of false negatives, for an individual class used in classification: i. If the classes are unbalanced - where there are unequal numbers of each object, but high performance in all classes is desired, it is often beneficial to judge performance using an average score over each class (Sánchez-Sáez et al., 2021).

\begin{equation}
Precision_{overall} = \frac{1}{n_{class}} \sum_{i=1}^{n_{class}} Precision_i
\end{equation}

Where $n_{class}$ is the total number of classes used in the model.

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>ecl</td>
<td>0.84</td>
<td>0.91</td>
<td>0.87</td>
<td>13238</td>
</tr>
<tr>
<td>lpv</td>
<td>0.93</td>
<td>0.94</td>
<td>0.94</td>
<td>2168</td>
</tr>
<tr>
<td>pul</td>
<td>0.86</td>
<td>0.77</td>
<td>0.81</td>
<td>9224</td>
</tr>
<tr>
<td>yso</td>
<td>0.86</td>
<td>0.73</td>
<td>0.79</td>
<td>1057</td>
</tr>
<tr>
<td>avg</td>
<td>0.87</td>
<td>0.84</td>
<td>0.85</td>
<td>25687</td>
</tr>
</tbody>
</table>

Table 3.2 The precision, recall and f1 corresponding to the above PRF, for each object type and overall.

The training results for the first part of the network are shown in Figure 3.21 and in Table 3.2 where the algorithm achieves an overall accuracy of 85.4%. Of note are the high precision and recall of LPVs and the lower accuracy of pulsating variables and young stellar objects - with the pulsating variables identified 20% of the time as eclipsing binaries. This is not a surprising result - contact binaries are populous in VIVA as mentioned earlier, and have a sinusoidal-like light curve similar to many pulsating variables.

For the second version of the network, as shown in Figure 3.22 the network does
Figure 3.21  Confusion matrix of the first section of the PRF - denoting the fraction of each object, by true label, and where they were assigned by the PRF. A value of 0.91 would correspond to 91% of all objects in that row assigned to the label corresponding to that column. In the example here, 12023 of the eclipsing variables were labelled as such, whilst 1215 were labelled incorrectly. The following abbreviations are used: ecl - Eclipsing Binary, pul - Pulsating variable, yso - Young Stellar Object.

Figure 3.22  Confusion matrix of the second section of the PRF.
not perform as well, with an overall accuracy of 73%. Of particular note is the overlap between the three LPV types, with all bar the OSARGs showing strong signs of overlap with the other classes. Difficulties in determining between the different types of LPV are noted in other works (Kim & Bailer-Jones, 2016), where there is much overlap between the OSARG and SRV and difficulty separating the two of them, resulting in accuracies of \( \sim 60\% \).

The accuracy of the OSARGs themselves is also peculiar. High value of accuracy is desired, but there will always be uncertainty and noise in any dataset and in any machine learning process there is the risk of overfitting. The network was trained on one part of the data and tested on a separate part, so it is unlikely to be an issue with the training.

However, with this peculiarity combined with the drop in performance splitting up the variability categories, using the first PRF and splitting up the LPVs by their period manually, seems to be a more promising option.

### 3.4 Convolutional Neural Network

The basic principle of the Convolutional Neural Networks (CNN) (LeCun et al., 1999) uses filters and convolutions that are trained to respond to different patterns in the image or input data. The network then learns the most critical underlying patterns and how they correspond to the provided classes, even quite precise features such as faces as part of face recognition, one of the CNN’s most employed functions.

In astronomical use Hon et al. (2018) trained a convolutional network on images of the spectra from RGB stars in order to ascertain their evolutionary states and Möller & de Boissière (2020) applied CNN’s to classify supernovae.

As in the above random forest approach, light curves are represented using statistical approximations or descriptors (Bloom & Richards, 2012; Nun et al., 2015). This is an expensive computational process that scales both with the number of objects, the number of measurements for each object and the number of descriptors. Upcoming surveys like the LSST will generate \(~330\) Mb/s for sustained periods of time (Zorich et al., 2020), necessitating high throughput transfer and processing. As science develops and surveys improve, this will lead to longer and longer processing time for any machine learning.
Additionally, these statistical descriptors cannot be reliably compared between different surveys - the light curves will differ intrinsically due to observational cadence or filter differences. Figure 3.23 shows how some common features such as the Skew can show noticeable differences between different original data sets.

This technique draws upon the works of Aguirre et al. (2019). The authors developed a novel deep learning model with which to classify light curves and avoid the above issues by "visually" studying the light curve and the differences between each point. The method can extract the information used for classification - in this case, the processed light curves - over 1,000x faster than statistical properties and variability indices can be calculated by feature analysis suites. Model accuracy per class between the CNN and the RF is comparable, with differences in the order of \(~0.01\) in the accuracy’s of each, with most of the model classes being tested at accuracy’s of \(~0.9\), or 90%.

### 3.4.1 Algorithms

As mentioned above, the light curve of each object is the basis for the "image" that the CNN will be "observing". Of the CNNs used for light curve classification, both Aguirre et al. (2019) and similar paper Mahabal et al. (2017) make use of the same technique. This technique involves mapping the light curves to $\Delta m - \Delta t$ space - representing each point by its difference from its chronological predecessor, both in time and in magnitude. This processed data becomes the input that the CNN sees.

Most convolutional networks use combinations of different layers in their architecture. The components of the CNN will be broken down into their component
functions and how they are used.

**First Convolutional Layer**

The convolutional layers are the key layers used in CNN. These layers extract information from their inputs using a sliding window with weights over the data, processing the information contained in the window with its weights with a non-linear function and then passing it on to the next layer.

The weights of the window are what the network optimizes. They are kept the same as the window moves over the input data, extracting information about the "local" region covered by the window. The improvements over the training improve the weights, better enabling it to extract specific information from the local data. The number of windows is set to the number of features the CNN should pick up over its training (Aguirre et al., 2019). The size of the window, its step size across the data and the number of windows used are hyperparameters, chosen before the model is trained and unchanging throughout.

The sliding process of the window can be shown in figure 3.24.

This convolutional layer is split in two parts - processing the magnitude and time data separately, but with the same weights. Keeping the weights this way is crucial - it enforces that the weighted windows the network learns are twinned and linked to both the period and magnitude of the object and that the features it is learning are drawn from both. The input to each layer takes 1D form, as is also demonstrated in figure 3.24. The number of windows was set to 64, with the value taken from the original paper by Aguirre et al. (2019). The value there comes from the number of features extracted from the light curves by feature analysis software (Nun et al., 2015) and as such, was used as the basis for what features the network could automatically extract using the convolution process.

**Second Convolution Layer**

A second convolutional layer is employed after the first. By using multiple layers, it increases the complexity of the features the network is capable of detecting (Jiang & Liang, 2016). This layer is set up in the same split time and magnitude approach as layer one. The number of filters in this layer is reduced to 32.
Figure 3.24  A step by step example of the convolutional process, demonstrating sliding the window or kernel across the data. The size of the kernel and the step size are hyperparameters, set before training. The values used will be discussed in section 3.4.2. Figure 5 from Aguirre et al. (2019)
Flatten Layer

This is the first of two fully connected layers. Fully connected layers are defined by having a connection between each node of this layer and the one they are connected to - in this case the next layer. This fully connected layer acts to merge the two parts of the convolutional layers as well as mix the information extracted (Jiang & Liang, 2016).

Hidden Layer

The hidden layer is the second of two fully connected layers, connecting back to the flatten layer and ahead to the softmax layer. The hidden layer is one of the most important. The nodes take in information from the flatten layer and use an activation function to respond in a non-linear fashion to them. This non-linear output allows the network to encapsulate yet more complex behavior.

Output Layer

This final layer contains one node for each possible output class of the classifier. The strongest signal across this node corresponds to the networks most confident prediction about a given object. The signals from these nodes are normalized (from 0 to 1) using a softmax function, allowing us to treat the resulting output like a probability of a light curve belonging to each node.

3.4.2 Implementation

The above network architecture was trained on a 6 core PC with a NVIDIA GeForce GTX 1080 graphics card. The CNN was implemented in Python 3 using the Keras framework (Chollet et al., 2015), with scikit-learn (Pedregosa et al., 2011) being used to assist in the training process and splitting the data into testing and training sets.

The following ranges of hyperparameters were tried.
Convolution Activation Function

3 different activation function for the convolutional layers were tested: Relu, Sigmoid, Tanh

The rectified linear activation function (Relu) is defined as $g(z) = \max(0, z)$, where $z$ is the input to the node. The Sigmoid activation function is $S(x) = \frac{1}{1+e^{-x}}$. The Tanh activation function uses the tanh function.

The Relu activation function performed massively superior here, with overall accuracy dropping by sin20%) if the Sigmoid and Tanh functions were used.

Fully Connected Activation Function

Relu has noticeably the worst performance (worse overall accuracy by ~5%). The difference between the Sigmoid and Tanh functions was very slight, but the Tanh format had better overall accuracy by ~1%.

Batch Size

The batch size denotes the number of training examples used in one pass down the network. Sizes of 256 and 128 were considered, and the larger batch size performed better.

Hidden Dimensions

I used 128 nodes for the hidden layer, taken directly from the paper of Aguirre et al. (2019). Values of 64 were tested, but no increase in performance was noted.

Dropout

A very successful way to improve model performance is to combine the predictions and labels of multiple models together, as has already been shown in the random forest used in Section 3.3. For a large and complex model, such as a CNN with multiple layers, this can be difficult and computationally resource intensive. Krizhevsky et al. (2012) suggest the use of a technique called “dropout”. By
setting some fraction of the hidden nodes to produce a flat zero output, these neurons are "dropped out" and do not contribute to the neurons in front of it.

This both reduces the size of the network and forces the network to learn more reliable features that are robust and useful even if a node before it is dropped.

Different values of dropout have been tested: 0.2, 0.3, 0.4 and 0.5, where 0.5 corresponds to a 50% chance of a mode being dropped. From testing at each value, 0.5 gives the highest performance, and has been chosen as such.

**Epochs**

This parameter defines how many iterations of successive training the model undergoes. Training for too long can lead to overfitting - where the model learns to fit the training data too well, effectively reproducing it. This leads to the model massively under-performing when exposed to new data, as it only learned the distinguishing properties for the training data.

To avoid this, 10 epochs of training are used, which provides a strong middle ground, giving good accuracy without evidence of overfitting.

**Kernel Size**

The kernel is the window that slides over the data and extracts information from each chunk it slides over. Too large a kernel size and the network fails to learn local patterns as well, but too small and it won’t see the patterns appearing. The original paper used a kernel size of 50 and after testing a kernel size of 40 and 60 with no improvement, a size of 50 was decided on.

**Validation Size**

The validation set was 25% of the total input objects. The values tested ranged between a 15% to 30% split, but without much noticeable difference. The value used for chosen simply as a midpoint.
3.4.3 Training

The number of objects available is not always equal - as was discussed earlier in the discussion of VIVA and OGLE, some classes of variable are more populous than others. This can be an issue. If one class in the overall dataset is much smaller than the remaining classes, there runs the risk of it’s defining characteristics not being picked up by the network, leading to failure in classifying such objects.

There are options for dealing with this, including:

- Permutations of existing light curves. An approach often taken with graphical image processing. It also helps to train the network at different configurations of the same object, such as an image upside down or a light curve starting at different points in phase.

- Generation of new light curves. If the light curve of the object is well understood, models of the object can be used to recreate a sample of new light curves from a range of given properties. This can be done for data not in light curve form – generating artificial images or extracting parameters such as temperatures from models.

This approach is difficult for our data. As discussed in Section 2.4, OSARGs show noticeable light curve differences in the NIR. Any modelling of light curves would have to be done from first principles, which there was insufficient time to do. In addition, the OSARGs used for such modelling of light curves are only drawn from OGLE – leaving all OSARGs that ogle was not able to detect unaccounted for in the modelling. Thankfully, training sample sizes were not an issue here.

The training data for this network were the Ks band light curves of VVV objects with variable type identified by crossmatching from OGLE. The final training sample consisted of 8000 of each of:

- Contact Binary
- Detached Binary
- OSARG
- SRV
Unlike the PRF, the CNN examines the light curves directly - performing a more complicated mathematical procedure with more nodes - and with restricting each light curve for each type to those with over 100 data points, the CNN must analyse over \( \sim 30 \) times more data points. The exact sample size of 8000 was used by the original creators of this CNN \cite{Aguirre:2019}. The fixed sample size has the advantage of not biasing the network by feeding it more of one type of object then any other. Testing with other sample sizes was considered, with decreasing the objects in each sample to 4000 worsening performance by a few percent as expected. Increasing the sample size above 8000 gave a marginal increase in performance but took longer to train and continued increase resulted in some of the samples not having enough candidates. To avoid this and with the benefits of a longer training mixed, a sample of 8000 for each was used.

### 3.4.4 Validation and Results

<table>
<thead>
<tr>
<th>Object Type</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact Binary</td>
<td>0.70</td>
<td>0.62</td>
<td>0.66</td>
<td>7448</td>
</tr>
<tr>
<td>Detached Binary</td>
<td>0.71</td>
<td>0.89</td>
<td>0.79</td>
<td>7793</td>
</tr>
<tr>
<td>OSARG</td>
<td>0.79</td>
<td>0.83</td>
<td>0.81</td>
<td>8461</td>
</tr>
<tr>
<td>SRV</td>
<td>0.69</td>
<td>0.67</td>
<td>0.67</td>
<td>7181</td>
</tr>
<tr>
<td>Mira</td>
<td>0.71</td>
<td>0.66</td>
<td>0.68</td>
<td>8075</td>
</tr>
<tr>
<td>Cepheid</td>
<td>0.85</td>
<td>0.77</td>
<td>0.81</td>
<td>7092</td>
</tr>
<tr>
<td>RR Lyrae</td>
<td>0.76</td>
<td>0.77</td>
<td>0.76</td>
<td>9950</td>
</tr>
<tr>
<td><strong>avg</strong></td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
<td>56000</td>
</tr>
</tbody>
</table>

**Table 3.3** The precision, recall and f1 corresponding to the CNN, for each object type and overall.

The training results for the first part of the network are shown in Figure 3.25 and in Table 3.3, where the algorithm achieves an overall accuracy of 74%. The OSARGs and Cepheids have the highest accuracy and are generally found best by the network.

The SRV and Mira are found with slightly less accuracy than the OSARGs, but most of the misclassification of the LPVs is with the other LPVs. Since all of
Figure 3.25  Confusion matrix of the CNN - denoting the amount of each object, by true label, and where they were assigned by the CNN. The following abbreviations are used (in addition to SRV, Mira and OSARG which have already been defined): Cep - Cepheid, RRlyr - RR Lyrae, ContactBin - Contact Eclipsing Binary, DetachedBin - Non-Contact Eclipsing Binary.
the LPVs have period-luminosity relations with overlap or close proximity in period-luminosity space \cite{ita2004}, these three types of objects should be separable using the periods from the period-finding process. The overlap between the LPV, other pulsating and eclipsing binary groups in the CNN’s predictions may indicate that the groups used to train have overlap within those groups - that is, some LPVs may be misclassified within the supergroup of LPVs.

The CNN overall is slightly less accurate (5%) across all objects than the PRF but I do not believe this will prove a large problem. With two methods that both produce satisfactory results, in much the same way that a random forest improves the performance of a single decision tree by compiling the results of multiple together, by using both networks on the same sample of objects and comparing their two predictions I can improve the confidence with which a classification is given.

3.5 Discussion and Conclusion

Collectively, the models can adequately identify LPV variables from a list of candidates, with the added prospect of an additional sample of eclipsing and pulsating variables as well, although the study of these will be left for future work or other researchers, due to time constraints.

There are improvements that could be made to the networks available as well as new networks altogether to consider. The choice of using the VIVA catalogue alone for the was not ideal as it ended up being fairly limited in terms of LPVs. The sample had no labelled OSARGs - in this case stars labelled as red giant branch (RGB) seemed to be the only previous detection’s of variability in that region of the HR diagram. It is possible that some of these objects were in fact SRVs or giant stars showing solar-like oscillations.

I have been developing a variant of the CNN used in Section 3.4. The version used here performed its convolution over the time and magnitude data in two separate “towers” with shared weights, I constructed a version with two time and two magnitude towers - one for VVV time and magnitude data, and one for the time and magnitude data from an optical survey. The network’s size proved too difficult to train on the hardware I had available to me, so this avenue has temporarily been put aside for now.
In comparison to the initial paper by Aguirre et al. (2019), the performance of my network is comparable to theirs for Cepheids and RR Lyrae. When comparing eclipsing binaries and the LPVs, the results vary as follows: My network outperforms the Aguirre et al. (2019) CNN for VVV eclipsing binaries by $\sim 15\%$ but underperforms for OGLE eclipsing binaries by the approximately the same margin. The Aguirre et al. (2019) CNN outperforms mine for both VVV and OGLE Mira by $\sim 20\%$ and OGLE SRVs and OSARGs by $\sim 10\%$ but is not set up to find SRV and OSARGs in the VVV data. My CNN outperforms the random forest used for comparison in the initial paper for SRV and OSARGs in the VVV data. I conclude that overall performance between the two CNN is mixed - their network performs very well on the OGLE data but less well or not at all on the VVV data whereas mine successfully extracts the OSARGs and other LPVs I am interested in but underperforms for the remaining variables.

My concerns with Aguirre et al. (2019) training the network on both near-IR (VVV) and optical (OGLE or Convection, Rotation and planetary Transit (CoRot)) has no clear conclusion. Initially I believed that due to the light curves of LPVs, Cepheids and RR Lyrae varying strongly between the two wavebands this would have a detrimental effect on training the network. Doing so has not negatively impacted their classification of OGLE data and my networks slightly improved VVV classification could be due to other factors - such as me using newer data with more epochs.

If this does prove to be an issue, it could be solved by training the network on optical and near-IR data side-by-side, so that it learns optical features from the optical data and near-IR features from the near-IR data, with the input data being objects with data at both wavelengths. Attempts at implementing this network failed due to hardware constraints - the additional convolutional layers and the need for larger hidden layers to better model the larger inputs were too taxing on the hardware I had.

Other approaches use networks such as Becker et al. (2020). They also attempt to avoid the time consuming process of feature extraction by using deep learning architecture based on Recurrent Neural Networks (RNN) (Lipton et al., 2015). RNN are specifically designed to work with sequential data, where each cell in the network has a memory that describes previous time sequence elements and uses that memory element in conjunction with each new data point.

In the Becker et al. (2020) paper, the performance of the network is compared to
that of a random forest on the OGLE, WISE and Gaia data. The performance of the method is comparable for determining broad superclasses of variables (LPV, eclipsing, Cepheid) but slightly worse at classifying at the subclass level (EA type binaries vs EB type binaries, for example). Thus, this method was considered but eventually not used.

To conclude, the networks each classify their samples with >75% accuracy with the PRF classifying at >85%. This accuracy will allow the classification of the final sample of variables with a <25% misclassification rate. By combining the predictions of two networks and finding the objects they agree on, I aim to reduce the misclassification rate and better extract the final sample of OSARGs and other LPVs to study using period data.
Chapter 4

Results and Application

With the networks successfully trained, what remains is to apply them to real and unclassified data - unseen to the trained networks. With the networks predictions the variable candidates can be labelled and any useful for specific studies can be selected from the whole.

There was not sufficient time to process all of the candidate sample - VIVA, with \(~45\) million objects. Indeed, since VIVA is designed to be almost complete for variable candidates, there will be objects falsely flagged. The number of true variables is more likely between one to ten million. As such, criteria must be devised to allow for application of the two machine learning networks to a selection of objects from the whole.

4.1 Final Candidate Selection

A smaller sample, consisting of the most certain objects, selected using the following cuts:

Using objects with many measurements allows for more accurate periods and more reliably calculated variability indices. The VIVA column vvvVivaCatalogue.nGoodMeasurements\(^1\) which counts the number of good measurements in the VVV DR4 - an observation with time, magnitude and magnitude errors, and no bad flags. These detection flags are a measure of observational quality and

\(^1\)When given in this form, this denotes the table and the column of that table in which this data is located in the Vista Science Archive.
label the presence of pixel issues or similar problems. An object was selected if it had over 80 good measurements in VVV DR4 and over 100 in VVV DR5. Other variability studies such as Kügler et al. (2015) used a threshold of 50 observations and Kim & Bailer-Jones (2016) used a threshold of 100 when creating an ultraclean sample. The values I used allow for selecting a sample of moderate size whilst also ensuring that the object has enough epochs of data to reduce uncertainty in any of its measurements.

The next cut used was to select objects with mean Ks band magnitudes (vvvVariability.ksMeanMag) above 10 mags - the threshold for which points would become saturated. In addition to this, all points in all objects with error flags that indicated saturation or pixel defects were removed. Rejecting only some points does bias the observations slightly towards to the fainter end - by cutting out information about bright objects that still managed to retain enough valid points from the first cut. Whilst there is still information that can be learned from saturated objects using these magnitudes in combination with the uncertain distances mentioned prior seemed like it would compound the errors in both sources.

Since this project has a focus on LPVs, which are generally cooler and redder, a selection was drawn on objects using J-K colour greater than 0 (J-K > 0). The object magnitudes used were vvvVivaCatalogue.jAperMag3 and vvvVivaCatalogue.ksAperMag3. Figure 2.1 shows examples of the sorts of colours expected for evolved low mass stars like Miras. OSARGs are similarly red, with Soszyński et al. (2007) finding most of their sample of OSARGs concentrated around (J-K) ~ 1.

A cut on the light curve amplitude in the VIVA table, vvvVivaCatalogue.aVar, was used. As mentioned above, this project has a focus on LPVs. As generally high amplitude stars, an cut of aVar > 0.1 was used to select all objects with amplitude above 0.1 mags.

Using the variability indices described in Section 3.2.4, I made cuts to the data based on information from Ferreira Lopes et al. (2020) and suggestions from the first author (private communication). They recommended using a combination of the power spectrum size vvvVivaCatalogue.HeightPKf2 and the false alarm probability vvvVivaCatalogue.faPcorrelation2 and note that a ratio of HeightPKf2/faPcorrelation2 > 1 was sufficient to select 95% of known crossmatched variables. However, the author also notes that some long period
objects such as Miras, as well as some aperiodic variables were not found. As such, I am using a threshold of $\text{HeightPKf2}/faP\text{correlation}^2 > 0.8$ - slightly lower than what they recommend to avoid missing any LPVs.

The authors (Ferreira Lopes et al., 2020) used their created X index (describing the ratio of $\sigma$ to its expected noise value $\eta$) to select known variables. This value was accessed at vvvVivaCatalogue.xIndex. They found that a threshold of $X > 1.5$ captured roughly 80% of objects and recommend using a cap of $X > 2$ to select the most reliable sample (Ferreira Lopes et al., 2020). This recommendation based on analysis of known objects is believed to be reliable and as such, $X > 1.5$ was used as the cut-off value here.

The final restriction on the objects used was that they could be crossmatched with Gaia in order for the PRF to have optical apparent magnitudes. This is a very strict restriction that limits the objects in our final sample to ones with well matched Gaia counterparts, requiring that the objects here be bright, not behind much dust and limited to the regions where VVV and Gaia have overlap.

### 4.2 Final Sample

Using the selection described above and including only those objects with parallax to allow some consideration of distance, resulted in a final sample of 189318 objects. The relevant information for the machine learning network was extracted as described in Chapter 3 and the list of processed candidates given as input to both networks.

#### 4.2.1 CNN Results

The above final sample, when used as input to the trained network resulted in a final catalogue shown in Table 4.1.

#### 4.2.2 PRF Results

The above final sample, when used as input to the trained network resulted in a final catalogue shown in Table 4.2.
### Table 4.1

<table>
<thead>
<tr>
<th>Classification</th>
<th>Counts</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact Binary</td>
<td>42811</td>
<td>22.6%</td>
</tr>
<tr>
<td>Detached Binary</td>
<td>3064</td>
<td>1.6%</td>
</tr>
<tr>
<td>OSARG</td>
<td>100661</td>
<td>53.1%</td>
</tr>
<tr>
<td>SRV</td>
<td>12042</td>
<td>6.3%</td>
</tr>
<tr>
<td>Mira</td>
<td>1330</td>
<td>0.7%</td>
</tr>
<tr>
<td>Cepheid</td>
<td>4306</td>
<td>2.2%</td>
</tr>
<tr>
<td>RR Lyrae</td>
<td>25104</td>
<td>13.2%</td>
</tr>
</tbody>
</table>

Table 4.1 **CNN Predictions of final sample of candidate variables.** Results are separated into the broader superclasses and variable subcategories. 
*Eclipsing includes Contact and Detached Binaries, LPV includes OSARGs, SRVs and Miras and Other Pulsating includes Cepheids and RR Lyrae.*

### Table 4.2

<table>
<thead>
<tr>
<th>Classification</th>
<th>Counts</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eclipsing</td>
<td>45875</td>
<td>24.2%</td>
</tr>
<tr>
<td>LPV</td>
<td>114033</td>
<td>60.2%</td>
</tr>
<tr>
<td>Other Pulsating</td>
<td>29410</td>
<td>15.5%</td>
</tr>
</tbody>
</table>

Table 4.2 **PRF Predictions of final sample of candidate variables.**

#### 4.2.3 Comparison and Other Works

Between the two methods most variable candidates are split into Eclipsing and LPV categories. The CNN classified more objects as LPVs and did so ~ 60% of the time - the largest single fraction of objects between the two networks. The PRF predicted a more even spread, with the exception of YSOs, which were predicted only a few handfuls of occasions.

The low YSO amount is not surprising - the final sample used included the cut of HeightPKf2/afPcorrelation2 > 0.8 - noted by the authors to reduce the numbers of aperiodic and semi-regular variables. YSOs fall in the former category.

Comparing the magnitude distribution of the two sets of predictions is shown in Figure 4.1. Of particular note is the very low magnitude for the LPV sample.
Figure 4.1  The $K_s$ band magnitudes violin plot for the two sets of predictions. Showing the results of the CNN (top) and the PRF (bottom). The graph denotes the distribution of the $K_s$ magnitude along the vertical axis.
produced by the PRF, which only goes to about 13 mag (compared to the CNN sample which reaches 17 mag). I believe that this may be due to the training data. The PRF was trained on a combination of VIVA and Gaia data and as such close objects with good matches were over-included. Filtering the training data to include a range of magnitudes was considered in Chapter 3 but is a tricky proposition which risks exchanging one bias for another. There is also the possibility that the other pulsating variables category includes some of the LPVs, since the two categories overlap in light curve morphology. Period is the strongest way to separate the two types, but I cannot be certain that every object has a reliable period. A further consideration would involve restricting both to the same sample. Much of the difference in shape is due to the different restrictions placed on the two datasets used as input and output for each - restrictions from Gaia crossmatching or large numbers of good epochs for the PRF and CNN respectively.

A comparison between the amplitudes of the different predicted categories is shown in Figure 4.2. High amplitudes were not expected and have only occurred very infrequently, as shown by the graph above. The overall similarity between the different types is unexpected however. The LPVs in each are concentrated at slightly above the lowest possible amplitude of 0.1 - the threshold used to initially select the sample, with the peak of the amplitudes in both cases being slightly higher than the other variable types. The pulsating variables being of comparable amplitudes to the LPVs is odd - especially in the near-IR where literature predicts much smaller amplitudes overall. This is perhaps another sign of some LPVs being falsely classified as pulsating variables.

The overlap between each prediction algorithm is shown in Figure 4.3. In comparing the two predictions, the two methods agree most strongly for the pulsating variables - 84.4% of the CNN pulsating variables matching the PRF and 63.1% vice-versa. The long-period variables match strongly in one direction but not the other - 42.5% of the CNN LPVs matching the PRF and 79.0% vice-versa. The eclipsing variables are responsible for the low matching of the above fraction. 74.1% of the eclipsing variables in the PRF have matches but only 38.4% of the eclipsing variables having matches in the CNN - the rest of them having been classified as LPV.

The overall mismatch is between the LPVs and the eclipsing binaries. One possible explanation may be because of the magnitudes of eclipsing binaries. Unlike pulsating variables, in eclipsing variables the change in light reaching
Figure 4.2 The Ks band amplitudes violin plot for the two sets of predictions, showing the variational amplitudes of the objects in each category. Showing the results of the CNN (top) and the PRF (bottom).
Comparing the predictions of the CNN and PRF, showing how each network classified each object and the amount of overlap or disagreement in their predictions. The most notable disagreement in the predictions is the PRF classifying $\sim 50\%$ of the CNN-predicted LPVs as eclipsing binaries - outnumbering the number of eclipsing binaries both networks agree on.
us is caused by the physical blocking of the light from the star. This means that the drop in light is (for our purposes) the same at all wavelengths, whilst pulsating variables are have smaller amplitudes in the near-IR (with the exception of OSARGs). This could lead to both LPVs and eclipsing binaries having similar amplitudes, confusing the networks. Arguments against this would be that eclipsing binaries can also have low amplitudes which should overlap with the pulsating variables if this were the case.

Overall, the agreement between the CNN and PRF for the pulsating and long-period variables is acceptable. The two networks agree most strongly about these objects, with the PRF predicting that many of the LPVs found by the CNN are actually eclipsing binaries. Overlap between these two could be due to the amplitudes of eclipsing binaries being generally independent of wavelength, resulting in amplitudes that can match those of LPVs, particularly as the LPVs amplitude falls at our higher wavelengths.

Whilst the CNN training data did not include any YSOs, the distribution of the PRFs predictions can still be considered. Three quarters of the YSO classified objects were predicted to be LPVs by the CNN. I believe the LPV are the most likely object to be classified as YSO because they are redder and can have unclear periodicity like SRVs stars (Rebull et al., 2014).

The recent paper by Molnar et al. (2022) describes their VIrac VAriable Classification Ensemble (VIVACE) catalogue, another attempt at studying variable stars in the VVV. They use a similar technique to ours in acquiring a final sample of variables - select a sample of likely variable candidates using simple statistical techniques and then employ a random forest to classify the likely candidates into variability types. Their approach differs from ours in two ways.

They use magnitudes obtained via the PSF photometry available in the VVV Infrared Astrometric Catalogue (VIRAC) databases (Smith et al. 2018), which are used to calculate simple variability statistics including the MAD and the Stetson I, J and K indices (as described in Section 3.2.4). This is in contrast to the indices used by this study, which include the indices created as part of the VIVA catalogue (Ferreira Lopes et al. 2020) and are based on aperture photometry. The VIRAC database contains 1364732 variable stars with their predicted labels.

When comparing the final sample of variables I used above with the VIVACE catalogue, crossmatching only returned ~ 5000 objects, far lower than expected.
Figure 4.4 The overlapping $Ks$ magnitude histograms for all VIVA objects, VIVA objects with Gaia counterparts and VIVACE. The vertical axis denotes the fraction of objects at each magnitude due to imbalances between the 3 catalogues.
The two datasets and their overlap were compared, with no little difference in sky distribution, amplitude or magnitude. The magnitudes of the two catalogues are compared in Figure 4.4. The VIVA sample trends towards slightly brighter objects and when crossmatched with Gaia the trend is amplified and more objects towards fainter magnitudes are lost. When using the whole VIVA catalogue crossmatched with Gaia DR3 instead of the trimmed sample ∼ 90,000 matches are found.

This was calculated using a narrow crossmatching radius of 1 arcsecond. I had some initial concerns that this would prevent some objects being matched to their counterparts, but tests showed that 75% of all VIVACE objects had VVV counterparts with well defined magnitudes, indicating that the low crossmatching radius is not the cause of this discrepancy. In short, the VIVACE catalogue matches true positives with the VVV catalogue well, but many of the objects they have detected do not show up in the VIVA catalogue. It is currently uncertain why this is the case. The trimmings made on the VIVA sample may be responsible but it is not clear which would cause a drop of this size. The amplitude cut of greater than 0.1 mag removes some objects but should leave the LPV and binary samples mostly intact.

The recent Gaia DR3 study on LPVs (Lebzelter et al., 2022) compares LPV candidates found in Gaia (using its general classification of variables module and its Specific Object Study module) (Rimoldini et al., prep) with the LPV catalogues produced by OGLE data releases 3 and 4. The Mira and SRVs, both in the C-rich, O-rich and unknown cases, are able to recoverable with > 80% minimum accuracy except for C-type Mira recovered at 70%. Similarly, OSARGs are recovered at 75%.

When only considering Gaia objects with periods, this number drops. The Mira recovery rate falls by ∼ 10% but the SRV recovery rate and the OSARG recovery rate falls by ∼ 40%, bringing the total returned OSARG fraction to 22%. As will be discussed in Section 4.3, obtaining periods for OSARGs was also more difficult than expected. It is possible that Gaia, which has a similar number of epochs to VVV and less than OGLE also struggled in this regard. As of writing this, access to the processing modules (and corresponding final dataset) responsible for Gaia DR3 LPV sample in unavailable, so the exact figures of the number of epochs Gaia used remains unknown. Lebzelter et al. (2022) discuss the filtering criteria used to select the LPV candidates and note that objects with fewer than 12 points were not included, a very low number of points to work with for a variable object. Another possibility is that the small optical amplitude of these
objects meant that they were not initially identified as variable by Gaia and as such do not have assigned periods.

When comparing the above LPV sample of 1720588 candidates (Lebzelter et al., 2022) with my labelled sample of 189318 objects (each of which has been matched to Gaia already to use their parallaxes), only 5836 LPVs from my sample had matches in the Gaia catalogue. So far two catalogues of LPVs - Gaia and VIVACE - both overlap very little with my VIVA sourced LPVs. There is a possibility that the correlated indices VIVA uses is causing it to identify objects as variable that the other two do not, but whether or not the identifications are accurate is uncertain as of this point.

4.3 Period-Luminosity Relations

As initially mentioned in Section 1.2.1 pulsating variable stars have relations describing the pulsation periods and brightness.

Some studies of PL relations for LPVs and their astronomical use have considered using the non-Mira LPVs before (Whitelock, 2013), but ruled them out - claiming that the presence of multiple sequences for SRVs and OSARGs would make distance estimation less certain if one was unable to determine which sequence an object lies on. To contrast, Mira variables only have one sequence, albeit with the possibility of binarity or long secondary periods to complicate matters. Other works argue that due to SRVs being more common than Mira due to being longer-lived that they make up for this drawback, especially as their seemingly-irregular pulsations get more understood (Trabucchi et al., 2021b).

The format of the Period-Luminosity relation has several variants. The bolometric magnitudes (Whitelock, 2013) are hardier to the effects of interstellar extinction, as are measures such as the Wesenheit index. Other surveys have used K band magnitudes directly - a single magnitude is simpler to measure and relatively good at piercing through dust. For my purposes, I will be using the Wesenheit index, for easier comparison with OGLE.

Other works have calculated the following relations. Chibueze et al. (2016) used the work, with LMC Miras, to calculate
Using the period-luminosity relation of Feast et al. (1989) and assuming a distance modulus of 18.6 mag (Whitelock & Feast, 2000).

Other literature values for period luminosity relations include the work of Whitelock et al. (2008) for Miras

\[ M_k = -3.51(\pm 0.20)(\log P) - 1.10(\pm 0.1) \]  

(4.2)

and for the semiregular variables (Knapp et al., 2003)

\[ M_k = -1.34(\pm 0.06)\log P - 4.5(\pm 0.35) \]  

(4.3)

More recent papers such as Sun et al. (2022) compare the K-band period-luminosity relations of a number of papers (Yuan et al., 2018, 2017; Iwanek et al., 2021a) as well as their own relations calculated using Gaia and VLBI parallaxes. They define a singular relation for all five sources:

\[ M_k = a(\log P - 2.30) - b \]  

(4.4)

where \(a\) and \(b\) are model coefficients found by each paper. Parameter \(a\) describes the slope of the relation, whereas parameter \(b\) describes the zero point. Their results are summarised in Table 4.3. The much higher scatter for the Milky Way relations is attributed to the higher relative scatter among the distances - the distances to the LMC and M33 objects are dominated by the distances to their host galaxies, allowing these objects to be quite effectively treated as if they all lie at the same distance.
<table>
<thead>
<tr>
<th>Mira Source</th>
<th>$a$</th>
<th>$b$</th>
<th>Scatter</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>LMC</td>
<td>$-3.77 \pm 0.07$</td>
<td>$-6.92 \pm 0.04$</td>
<td>0.12</td>
<td>(Yuan et al., 2017)</td>
</tr>
<tr>
<td>M 33</td>
<td>$-3.77$</td>
<td>$-6.97 \pm 0.01$</td>
<td>0.21</td>
<td>(Yuan et al., 2018)</td>
</tr>
<tr>
<td>LMC</td>
<td>$-3.30 \pm 0.46$</td>
<td>$-6.67 \pm 0.06$</td>
<td>0.27</td>
<td>(Iwanek et al., 2021a)</td>
</tr>
<tr>
<td>Milky Way</td>
<td>$-3.59 \pm 0.29$</td>
<td>$-6.79 \pm 0.15$</td>
<td>0.45</td>
<td>(Sun et al., 2022), VLBI parallaxes</td>
</tr>
<tr>
<td>Milky Way</td>
<td>$-3.63 \pm 0.30$</td>
<td>$-7.08 \pm 0.29$</td>
<td>0.94</td>
<td>(Sun et al., 2022), Gaia parallaxes</td>
</tr>
</tbody>
</table>

Table 4.3  A comparison of Slope ($a$) and Zero-point ($b$) for a collection of Miras in the local group.

Figure 4.5  The Period-Absolute $K$s band magnitude plot for the sample of LPVs found by the CNN, as described above. The red line shows the location of the Sun et al. (2022) period-luminosity relation for Mira, as described in Section 4.3. The period here is found using the Lomb-Scargle periodogram.
4.3.1 PLR for VVV objects

The sample of variables can be used to derive period-luminosity relations as described above. More well researched variable stars including Cepheids, RR Lyrae and Mira have already been studied and some knowledge about them already exists. OSARGs are less well understood. Figure 4.5 demonstrates the period-absolute magnitude space for the final sample of objects used in this project. There are several key issues immediately obvious. There is very little indication of any horizontal or diagonal structure that would be indicative of any PLR.

The vertical lines seemed to be a sign of artifactual and the generation of false periods, but when zooming in on this region as shown in Figure 4.6, the periods are distributed in clumps. The breadth of the first region is approximately 38 days and the second is 48 days. The first is centered on 338 days and 416 days. I am not certain what observing cycles these correspond to and the clumps seem broader than expected for artifactual objects. There are other clumps at shorter periods including one at 79 days and the densest clump at 6 days.

The initial hypothesis for the identity of the first clumps in Figure 4.6 is that it represented an aliased period of 1 year. This signal would be expected, but raises some questions. Firstly, the clump is broader than expected (Baluev, 2008) for such a strong signal as well as not being centered on the year mark in question. Secondly, there is a slightly less dense adjacent clump of objects with similar properties. If one of these represents a year I cannot provide an explanation for the second clump.

Both the LSG and QMI periodograms produced similar results which causes me to believe that it is a problem with a shortage of observations rather than issues with periodogram frequency grid, for example. The frequency grid was set with a maximum and minimum frequency of 10.0 and 0.0001.

A comparison with the original crossmatched sample of VIVA is shown in Figure 4.7. They exhibit a strong trend of short (<1 day) periods and very few objects in the period range expected for OSARGs. Of note is the vertical line of points at an aliased period of 1 day. This structure occurred even when the VIVA catalogues strictest period-aliasing filters were used, which removed all occurrences of multiple objects having the same period - which only should occur for aliased periods.
Figure 4.6 As Figure 4.5, but zoomed in on the seemingly structured regions of period - the vertical lines.
Figure 4.7  *The Period-Wesenheit relation for the original sample of literature-crossmatched OSARG stars in VIVA. A logarithmic axis has been used to include the several high period objects.*

Other studies of LPVs in the galaxy have had mixed success. Quiroga-Nuñez et al. (2020) plot their sample of local LPVs with periods available in Gaia and compare and contrast with the predicted PL relationships from literature in Figure 4.8. The PL relationships used in the figure were calculated in the LMC and corrected for a distance modulus of 18.49 mags. Of note in this figure is the large cluster of objects that lie between the C and D sequences and there is subsection of points that lie on the D sequence. The authors note that the Whitelock et al. (2008) and Matsunaga & IRSF/SIRIUS Team (2007) relations were calculated using LMC objects and is biased towards the most luminous objects, whereas their sample is specifically targeting the closer, less luminous objects. They cite this as a possible explanation for the mismatch - the LMC PL relations were calculated without the large clump occurring in their data.

Overall, the final collection of variable stars have patches of noticeable disagreement, but the remaining data has a rough agreement of ~75%. The lack of any successful periodogram is disappointing but with more epochs I believe the periodograms to be improved to provide useful information about the periodic nature of these LPVs.
Figure 4.8  The Period-Absolute K band magnitude relation for the sample of local variables from the Bulge Asymmetries and Dynamical Evolution (BAaDE) project (Quiroga-Núñez et al., 2020). The absolute K band magnitudes are calculated using distances acquired from Gaia parallaxes. The literature LPV sequences from Whitelock et al. (2008) and Matsunaga & IRSF/SIRIUS Team (2007) are solid and dashed lines respectively and the sequences with letters corresponding to different pulsation modes (Lebzelter et al., 2019). Modes C and C’ are fundamental and first overtone Mira and mode D denotes long secondary period objects. Figure 13 from Quiroga-Núñez et al. (2020)
Chapter 5

Conclusion

In review, I have created a catalogue of classified variable stars from the larger catalogue of VIVA variable candidates and observed unexpected high amplitude near-IR variability from OSARGs, the cause of which still remains unknown.

Some modelling was attempted to understand the \( \sim 5x \) greater amplitudes of the OSARGs in Ks band as opposed to the OGLE I band. Other pulsating variable stars show the opposite behavior - greater amplitudes at shorter wavelengths (Fraser et al., 2008). Model spectra from X-Shooter (Verro et al., 2021) were used to mimic real spectra of RGB and AGB and objects with similar temperatures and the same masses and metallicities were compared to model the variation that occurs in an OSARG. This approach did not provide a clear explanation of the discrepancy, possibly because the effect is due to circumstellar envelopes of the OSARGs, which were not included in the spectra models.

Two trained machine learning networks were trained on variable stars from OGLE (Udalski et al., 2015), Simbad (Wenger et al., 2000) and AAVSO (Henden et al., 2016) with \( \sim 75% \) and \( \sim 85% \) accuracy for the Convolutional Neural Network and the Probabilistic Random Forest respectively. The predictions of these two networks combined were used to select the final catalogue of classified variables.

The periodogram analysis of said objects did not return periods that made sense for their classification and many seemed to show signs artifacting in the period finding. Without accurate periods and with some trouble establishing distances to objects in order to calibrate a period-luminosity relation, I was unable to establish a period-luminosity relationship for OSARGs in the Ks band. Likewise
the period-luminosity relations for SRVs and Miras could not be found.

Despite this, a large catalogue of variables was produced (189318 in total) with plans to expand this number to include more of the VIVA catalogue - increasing completeness with a potential drop in reliability. This sample was classified using two different networks and for objects where the two networks agree on a classification, I believe that these objects are the most reliable from the overall list to conclude are truly variable and the variable predicted.

This most certain sample - stars that both networks agree on a single classification - contains ~ 34,000 eclipsing binaries, ~ 48,000 long-period variables and ~ 25,000 other pulsating variables.

Overall, several questions about the OSARGs and their properties (amplitudes and period-luminosity relations) still remain, but for the amplitudes I believe I have ruled out systematic effects and identified a probable cause. A large sample of OSARGs will aid future work by myself and others, especially if additional data for better period finding or improved period finding techniques can be used.

When considering the processing of dust in future work the extinction law used may need to be reconsidered. For this work I used the extinction law for a diffuse medium [Voshchinnikov, 2012], which I considered considered an acceptable approximation due to the reliance on auxiliary information provided by optical surveys for distances and variable star classifications. When working outside of the regions clearly visible to the optical surveys, an IR-based extinction law such as the proposed extinction law of Fritz et al. (2011) may be more applicable. The paper by Fritz et al. (2011) suggests that there might be as great as a ~1 mag in Ks between their new model and that of the Cardelli et al. (1989) extinction law. This sizable difference would be worth investigated further.

In retrospect and upon further examination, the amplitudes for the LPV magnitudes are dominated by the uncertainty on individual epochs. This calls into question how accurate any amplitude measurements are and may have impacted classification using the PRF as the amplitude is a strong way to distinguish between the different types of LPVs. It will not impact the results of the CNN as this architecture does not use the calculated amplitude and would ignore any artifacts arising from variable indices.

Future research that can be carried out on this topic has been split into topics that could be carried out using the data already available and topics that would
require the collection of new data.

5.1 Future work using the same data

The first priority for future research is to expand the small sample to a large batch of objects. As discussed earlier, the cuts made to the VIVA sample which were designed were very strict and I believe may have removed some true variables. There is always going to be a trade-off between completeness and reliability, but I believe the cuts made may have caused too harsh a drop in completeness in exchange for only a marginal increase in reliability. A revised and broader set of criteria will be used to better capture all the variable objects in VIVA with more time available to process the larger batch of data. An additional avenue of work would be to apply the two networks to all 45 million VIVA objects. This would be time consuming, but having the predictions for all objects - including those unlikely to be variable, would help in distinguishing any false positives that occur in the improved sample of VIVA objects. By incorporating non-variable objects into the machine learning process, the networks would be better able to handle any non-variable objects as well as letting us make use of their probability based outputs to predict the likelihood of an object being variable.

In Chapter 3 there are two peculiarities I observed with my data. The first involves the use of Gaia parallaxes for distances to our LPVs produces distances I believe not correct. The size and varying photocenter of the LPVs leads to inaccurate parallaxes, a known phenomenon in the literature (Chiavassa et al., 2018; Sudou et al., 2019). This is one explanation, but the scale may be too large for the parallax errors occurring. The angular diameter of AGB stars can reach multiple milliarcseconds (Xu et al., 2019), far greater than the 0.3 milliarcseconds parallax errors observed from the OSARG sample. This issue may be compounded with the prior being calculated for non-bulge and disk regions of the sky.

In previous works this has mostly been applied to AGB stars such as Miras which leaves me uncertain if/to what extent we see parallax issues in less evolved LPVs like the OSARGs or if there are other systematic issues involved. If other LPVs, particularly OSARGs, could be sourced with very reliable distances then a thorough comparison could be made between their predicted parallax distances and their distances calculated via other means.
Etoka & Engels (2022) suggest a non parallax based method of distance estimation. They use the strong OH maser emission produced in the circumstellar envelope of evolved intermediate mass stars to track the motion and phase-lag of the envelope. Using a combination of single-dish and interferometric monitoring, the red- and blue-shifted maser emissions can be used to infer the distance to the object. This would require new data spectroscopic data, but does demonstrate how the existence of techniques to circumvent issues with one tool for probing an object. Using this technique requires a radio telescope and telescopes with interferometric capability’s to study the stars surface. The Very Large Telescope Interferometer (VLTI) or Atacama Large Millimeter/submillimeter Array (ALMA) could be used to collect the interferometric data needed. MIDI, an older decommissioned instrument used to probe AGB stars was able to reach K~7 objects with 1 hour time slots and GRAVITY (Gravity Collaboration et al., 2017), a newer instrument can reach K~10 for single field with 1 hour on a target with high precision. It may be more efficient to find closer OSARGs using survey such as 2MASS or Gaia in order to require less time on a single object.

The second peculiarity I noticed in Chapter 3 involved the discrepancy in absolute magnitude for the sample of LPVs studied. The sample as a whole was ~ 2 mag fainter than expected. I have no current hypothesis as to why this is occurring, but the above investigation into the distances to these objects would help to confirm or rule out bad distances as a contributing factor. As it stands, an individual object or cluster of objects could lie at the predicted upper limit of Gaia distances - ~ 10 – 20kpc and underestimating the distances to this degree would explain their faintness, but much less likely to be true for the numbers of objects it is occurring on. This could be caused by extinction, but seems unlikely as it would require several magnitudes of extinction in Ks across many objects to explain the overall shift of the population considered in that chapter.

The requirement for the above sample to require Gaia parallaxes in order to calculate the absolute magnitude is also a restriction that I believe impacts how complete the fraction of redder and dustier objects processed was. Future versions of this work will investigate replacements for the absolute magnitude for the PRF and whether or not they are necessary for object classification.

Further improvements to the PRF could improve on use of errors for each feature. Some of the simpler features such as magnitude were simple to incorporate errors into but other correlation based indices such as $K^2_{Fi}$ proved too difficult. In future work Monte Carlo error propagation could be used to estimate errors and improve
the performance of the PRF using these known errors.

With the improved and expanded sample of variables, my catalogue of classified variables for both methods, the accuracy of each and how the two compare for each object will be made publicly available. The Vista Science Archive and Vizier are the planned distribution methods for the data.

5.2 Future work using new data

To further understand LPVs - and in particular OSARGs, I think focused spectra of a handful of OSARGs could prove scientifically valuable. In the same vein, high spatial resolution observations give us a better understanding of the surface - and from that the composition of the circumstellar envelope and how it interacts or absorbs the emission from its star.

A recent example of the latter is the study of Betelgeuse during its recent dimming by Montargès et al. (2021). Even though it is a bright and close (220 parsecs away) star, the focused and higher resolution observations strongly helped in detecting the activity in the dust surrounding the star.

Ideally for both of these sets of new information, they would include repeated observations taken frequently enough to capture the variability cycle of the OSARG. This would give us spectra and/or high resolution images of its surface throughout every stage. I believe that this would be the best way of understanding why the amplitudes of the OSARGs are greater in the near-IR than the optical as was expected, by using snapshots at each stage of the pulsation to better grasp if the object has a circumstellar envelope that must be accounted for or any other not detected phenomenon.

The Apache Point Observatory Galactic Evolution Experiment (APOGEE) Survey (Allende Prieto et al., 2008) could be a facility to use for this. The survey and its successor APOGEE-2 have finished and feature near infrared spectra - between 1510 and 1700 nm, approximately H band and resolution of $R \sim 22,500$ of red giant stars in the Milky Way. APOGEE-2 includes the repeated observations of some objects required. Additionally, it covers the bulge region between $340 \leq l \leq 20$ and $|b| \leq 25$, giving some overlap with VVV and my existing sample of OSARGs.
Another option is the future spectrographic project built for the VLT is the Multi-Object Optical and Near-infrared Spectrograph (MOONS) \cite{Cirasuolo2020}. MOONS has a key advantage over APOGEE - it observes both optical and near-IR which I think would allow for a more detailed understanding of OSARG circumstellar dynamics than near-IR alone.

Another VLT instrument - GRAVITY - would be very useful for the high resolution images \cite{GravityCollaboration2017} whilst also giving extremely accurate astrometric detail in the K band, hopefully avoiding the issues with optical parallax observed via Gaia.

In conclusion, although the target goal of Period-Luminosity relations for OSARGs were unable to be measured, we will soon release a still expanding catalogue of OSARGs for other works, as well as plans to further investigate their unexpected behaviour.
Bibliography

Allende Prieto C., et al., 2008, Astronomische Nachrichten, 329, 1018
Astropy Collaboration et al., 2018, AJ, 156, 123
Avvakumova E. A., Malkov O. Y., Kniazev A. Y., 2013, Astronomische Nachrichten, 334, 860


Breiman L., 2001a, Machine Learning, 45, 5

Breiman L., 2001b, Machine Learning, 45, 5


186


Ferreira Lopes C. E., et al., 2020, *MNRAS* 496, 1730


Freund Y., Schapire R. E., 1996, Experiments with a New Boosting Algorithm


Fruth T., et al., 2012, *AJ* 143, 140

Hubble E., 1929, Proceedings of the National Academy of Science, 15, 168
Jayasinghe T., et al., 2018, MNRAS.
Jeffery C. S., 2008, Communications in Asteroseismology, 157, 240
Macri L. M., Ngeow C.-C., Kanbur S. M., Mahzooni S., Smitka M. T., 2015, AJ 149

194


Merc J., Gális R., Wolf M., Leedjärv L., Teyssier F., 2019, Open European Journal on Variable Stars, 197, 23


Minniti D., et al., 2010, New Astron., 15, 433


Raghavan D., et al., 2010, [ApJS] 190, 1


Reimann J. D., 1994, PhD thesis, UNIVERSITY OF CALIFORNIA, BERKELEY.

Reis I., Baron D., Shahaf S., 2019, [AJ] 157, 16


Rimoldini et al. 2022 (in prep), A&A

Rivinius T., Carciofi A. C., Martayan C., 2013, Astronomy and Astrophysics Review, 21, 69


Rodrigo C., Solano E., 2020, in XIV.0 Scientific Meeting (virtual) of the Spanish Astronomical Society. p. 182


Samus’ N. N., Kazarovets E. V., Durlevich O. V., Kireeva N. N., Pastukhova E. N., 2017, Astronomy Reports, 61, 80
Sana H., et al., 2012, Science, 337, 444
Sen S., Agarwal S., Chakraborty P., Singh K. P., 2022, Experimental Astronomy, 53, 1


Soszyński I., et al., 2007, Acta Astron., 57, 201


Soszyński I., et al., 2011a, Acta Astron., 61, 217


Soszyński I., et al., 2013a, Acta Astron., 63, 21


Sterken C., Jaschek C., 2005, Light Curves of Variable Stars


Sudou H., et al., 2019, PASJ, 71, 16


Taniguchi D., Yamazaki K., Uno S., 2022, Nature Astronomy, 6, 930

Taranova O. G., Shenavrin V. I., 2017, Astronomy Reports, 61, 983
VISTA/ESO, VISTA FoV Comparison, https://www.eso.org/public/unitedkingdom/images/vista_fov_comparison/
Virtanen P., et al., 2020, Nature Methods, 17, 261
Wallis J., 1693, Philosophical Transactions of the Royal Society of London Series I, 17, 584