Exploring the Infrared Variable Sky with Machine Learning

Supervised by:
Prof. Philip W. Lucas
SECOND SUPERVISOR
Dr. Yi Sun

Author:
Niall MILLER

Centre for Astrophysics Research
Physics, Engineering and Computer Science
University of Hertfordshire

Submitted to the University of Hertfordshire in partial fulfilment of the requirements of the degree of Doctor of Philosophy.

March 2024

Abstract

The last decade in astronomy has seen the growth of time-series data and with it, the emergence of large surveys. Surveys such as PTF (Law et al., 2009), ZTF (Bellm et al., 2019), CoRoT (Auvergne et al., 2009), HOYS (Froebrich et al., 2018) and VVV (Minniti et al., 2010) provide large amounts of large-area, multi-epoch data. Such surveys bring a multitude of new issues, many of which are in the form of 'unknown-unknowns'. From this, novel techniques are required to properly analyse these data. Manual analysis is unfeasible and hence, efforts have been taken to develop tools that seek to automate large portions of the data analysis. The new dimension of study afforded to us by these surveys allows us to probe the formation, evolution and death of stars in unique ways.

A fundamental issue arises **"How can we completely and robustly extract information from modern astronomical time series data?"** – Answering this question requires the development of novel methods and the improvement of those already established. In doing so, I aim to further expand and explain the demographics of variable stars in the Milky Way. By coupling more sensitive and robust identification methods with more thorough and complete analysis, I aim to identify and characterise new and known stellar classes. These actions seek to provide a more complete and accurate view of the Milky Way, its structure and demographics.

Key contributions of this thesis include the development of a neural network-based false alarm probability (NN FAP) method, which significantly improves the identification of periodic variables in large-scale surveys like VVV, LSST, and TESS. This method generates a universally comparable and unbiased FAP, making it applicable across various types of variable stars, leading to a more complete view of the demographics of periodic variable stars. The creation of the PeRiodic Infrared Milky-way VVV Star-catalogue (PRIMVS) underscores the effort to identify periodic variable stars comprehensively and without bias. Utilising the VVV survey's depth and breadth, PRIMVS processed over 86 million candidate variable sources using multiple period-finding methods and a novel neural network-based false alarm probability, leading to the identification of approximately 5 million periodic variables. Moreover, the thesis introduces a contrastive learning approach based on the SimCLR framework with a gated recurrent neural network (GRU) backbone, specifically designed to handle stochastically sampled time-series data. This method improves variable star classification by creating semantically meaningful embeddings, enabling more nuanced and accurate analysis. Additionally, the integration of VVV data with Gaia astrometry enhances distance measurements to star forming regions, while the use of Denoising Diffusion Probabilistic Models (DDPMs) for generating synthetic light curves provides a novel solution for developing extensive training sets.

Declaration

I declare that no part of this work is being submitted concurrently for another award of the University or any other awarding body or institution. This thesis contains a substantial body of work that has not previously been submitted successfully for an award of the University or any other awarding body or institution.

The following parts of this submission have been published previously and/or undertaken as part of a previous degree or research programme:

1. Chapter 2 - *The verification of periodicity with the use of recurrent neural networks*: This has been submitted for publication in the *The Royal Astronomical Society Techniques and Instruments* journal.
2. Chapter 4 - *Contrastive Curves*: This chapter builds on some preliminary collaborative work with Dr. Mike Smith.

Except where indicated otherwise in the submission, the submission is my own work and has not previously been submitted successfully for any award.

Acknowledgements

This work was only possible by standing on the shoulders of the giants that have taught me. I would like to extend great thanks to my PhD supervisor, Phil Lucas, for having the faith in me to complete this PhD. I hope to attain similar heights in my academic career and thank Phil for the standards he has shown me to hold myself to. The same is true for my other supervisor, Yi Sun. Thank you for providing a voice of clarity in the incredibly fast and chaotic world of machine learning.

I extend my gratitude to all of the peers and colleagues, in particular; Marina, Matt, Kasia, Mike K, Mike S, Rob, Vito.

Calum, my Dr. Brother, we have had many many many conversations in 2E50. Thank you, I enjoyed all of them. Be that the 5% of astronomy related discussion, the 50% of palaeontology related discussion or the rest that were mostly about `redacted` and `redacted`.

Will, I thank you for your active efforts in keeping me sane and cultivating a perfect vibe of scum and villainy in our abode, Rancho Relaxo, Millers rise.

Id like to thank Zhen Guo who has helped me far more than I ever could have expected. A lot of my confidence and networking is a product of Zhen's help and I look forward to a career in astronomy working with you.

Id like to extend thanks to my long friend and handler Rowan. You have immeasurably helped me with your genuine interest in my research and constant assistance with proof reading and sanity checking.

My academic career as a whole would not even remotely have been possible without the endless support from my family. I would have been destitute by many facets if not for my family and I will not forget their help. In particular, I dedicate this thesis to my late Grandma, Margaret Miller. Margaret nourished my love for science since I was very small and I am proud of my self for doing well for her and all of my family.

To Katherine, meeting you was possible through the adventures that are offered to a fledgling astronomer. I am forever thankful for your presence in my life, you have helped me incredibly with this thesis and everything else. You remind me that, while I love astronomy, it isn't the sole or foremost love in my life.

Contents

List of Figures

List of Tables

- *“When a distinguished but elderly scientist states that something is possible, he is almost certainly right. When he states that something is impossible, he is very probably wrong.”*
- *“The only way of discovering the limits of the possible is to venture a little way past them into the impossible.”*
- *“Any sufficiently advanced technology is indistinguishable from magic.”*

Profiles of the Future
Arthur C. Clarke

Chapter 1

Introduction

Variable stars are stars whose brightness as seen from Earth fluctuates over time. These fluctuations can be periodic, quasi-periodic, or aperiodic, each of which provides insights into the physical processes occurring within the star or in its surrounding environment. Periodic variability occurs when a star's brightness changes at regular intervals, quasi-periodic variability occurs with irregular intervals, and aperiodic variability occurs without any predictable pattern. Understanding the origin and nature of these variations is crucial for mapping the structural components of our galaxy, studying stellar evolution, and identifying unique celestial phenomena.

This thesis shows the tools and methods that were developed to most completely identify and analyse variable stars found in the VVV survey. There is a focus on periodic variable stars although quasi- and aperiodic variable stars are also analysed, albeit as a byproduct of periodic variable analysis.

The VVV (VISTA Variables in the Vía Láctea) survey is a large-scale infrared survey aimed at studying the Milky Way's bulge and disk. Its primary objective is to detect and catalogue variable stars to gain insights into the structure and evolution of our Galaxy. Unlike many optical surveys, VVV's infrared capabilities allow it to see significantly deeper through interstellar dust. Infrared wavelengths can penetrate dust clouds more effectively than optical wavelengths, allowing the study of Galactic regions that are otherwise obscured. This can uncover stars hidden from view in the optical range.

Due to the amount of time-series data produced by the VVV survey (800 million unique sources over 560 deg^2 in the VVV Infrared Astrometric Catalogue (VIRAC2b) Smith et al. (2018)) we must rely on robust and versatile methods to automatically extract reliable and accurate features.

Time-domain astronomy focuses on observing astronomical sources over time to detect changes in their brightness, position, or other time-dependent properties. The advent of time-domain astronomy has revolutionised our understanding of the universe by revealing variable stars, from which many physical properties can be derived through the analysis of their variability. Transient events, such as various types of novae, supernovae, gamma-ray bursts, tidal disruption events, active galactic nuclei, microlensing events are also identified and examined wihin this field. This field has also enhanced our understanding of stellar evolution, exoplanet transits, and the dynamic processes occurring in the interstellar medium.

As variable sources can change their brightness for a wealth of intrinsic and extrinsic reasons, significant efforts must be given to ensure our methods are robust enough to deal with the variety of variability present in the night sky. The following questions were answered or explored during this thesis;

What tools can we develop and use to more completely extract variable star information from the VVV survey?
The uniqueness of the VVV data set brings with it problems pertaining to unknown unknowns. We do not have a complete understanding of the expected types of infrared variable stars throughout the Milky Way. We can use existing wide-field temporal surveys such as Gaia (Gaia Collaboration et al., 2018) or Zwicky Transient Facility (ZTF) (Bellm et al., 2019) but these have different features to VVV (such as wavelength regimes or photometric and temporal depth) and hence different selection biases. With these differences recognised, efforts have been made to ensure that methods used for analysis are the most robust and impart the least biases possible. This was achieved with the development of novel tools, particularly with the use of Machine Learning. Chapter 2 discusses how a novel machine-learning technique was developed to combat the unreliability of classical techniques for the identification of periodic variable stars. Chapter 3 discusses the bespoke pipeline that was constructed to reliably analyse the time series data from VVV. Chapter 3.8 discusses a method to classify light curves by the analysis of the phase folded light curve itself.

What are some of the known and unknown populations and demographics of the infrared variable sky?
Once I have identified and constructed the most appropriate tools, I can more confidently explore the features of any constructed catalogue or data set. The earlier steps in this process will be most easily achieved by comparing to known and expected demographics. We can do this by cross-matching, classifying and number counting. Cross-matching is a 5-minute exercise in Tool for OPerations on Catalogues And Tables (TOPCAT)(Taylor, 2005) and this has already been done with the likes of Gaia, INT Photometric H-Alpha Survey (IPHAS)(Corradi et al., 2008) and Wide-field Infrared Survey Explorer (WISE)(Wright et al., 2010). Classifying the objects is a difficult task if the size of the survey precludes manual inspection. This will be achieved with machine learning techniques such as those illustrated in section 3.8 where the main goal is large population identification (achieved via the clustering of latent space). The identification

of large groups involves tuning latent space projections along with clustering techniques. There was also a brief look into using dynamic time warping as a way of classification.

Separate from the analysis of populations of objects we can look for individual, or smaller groups of, exotic sources. Effectively a more targeted and deep analysis of objects. The first step in this process is their identification. This is achieved with the use of the tools already developed. For instance, this could be dimensionality reduction of the catalogue with selected features that are astrophysically descriptive (period, amplitude, colour...). The identification of the more exotic objects is an extension of the methods used to identify larger more known groups.

How else can we utilise novel methods with the VVV survey?
The uniqueness of the VVV survey brings with it many new applications of well understood methods. We can also utilise machine learning to increase the combined effectiveness of multiple surveys. Chapter 3 section 3.7.3 shows an initial look into the use of external catalogues for the classification of VVV light curves. Chapter 5 section 5.1 shows work on the use of Gaia optical astrometry with VVV infrared astrometry to create more reliable star forming region distances.

1.1 Variable Stars

Variable stars can be categorised according to their underlying causes of brightness changes. Intrinsic variables experience changes due to processes within the star itself, such as pulsations or eruptions. Extrinsic variables, on the other hand, undergo variability due to external factors, such as eclipsing binary systems where one star passes in front of another.

The properties of an object's variability can change significantly depending on the source of variability. For example, W Ursae Majoris (W UMa) variables are a relatively common low mass contact binary of typically F, G or K type stars. W UMas generally have an orbital period of less than a day with an observed amplitude (Δm) of a few tenths of a mag. FU Orionis type stars are pre-main-sequence stars commonly found in star-forming regions, which have displayed an extreme change in their magnitude and spectral type, known as FUor events (Herbig, 1966). Typically, this change happens over approximately 1 year and is expected to last on the order of decades. It is clear that when designing a time-domain survey, aspects of the survey will change based on the scientific goals of the survey. It follows that time-series astronomical surveys and the light curves they produce can vary significantly in their sampling rates, temporal coverage, surveyed region, wavelength coverage, and quality.

The result of surveys with such significant differences means that a direct comparison between light curves of the same source from different surveys can often be difficult. Hence, to combat this, a source's light curve is analysed and measurable heuristic properties of the light curve are extracted.

The General Catalogue of Variable Stars (GCVS) (Samus' et al., 2017) classifies variable stars into several groups based on the mechanisms driving their variability. This classification encompasses a wide range of stellar phenomena from intrinsic changes in the stars themselves to extrinsic factors that alter the light we receive from them. An overview of these groups is visualised in a classification tree (Figure 1.1), which is adapted from the work of Eyer and Mowlavi (Eyer and Mowlavi, 2008).

Intrinsic variables include stars that change in brightness due to the physical mechanisms internal to the star. Eruptive variables, for instance, show variations in luminosity due to explosive processes or flares on their surfaces. Rotating variables change brightness as they spin, due to surface inhomogeneities like star spots or magnetic fields. The complexity of intrinsic variability reflects the depth and complexity of the physics governing stellar interiors and atmospheres.

Extrinsic variables, in contrast, owe their variability to the line-of-sight effects. This includes the light from a star being blocked by another star in an eclipsing binary system, or the gravitational microlensing effect of a massive object passing between a star and the observer. The diversity of extrinsic variables provides insights into the dynamics of binary systems and the structure of the Galaxy.

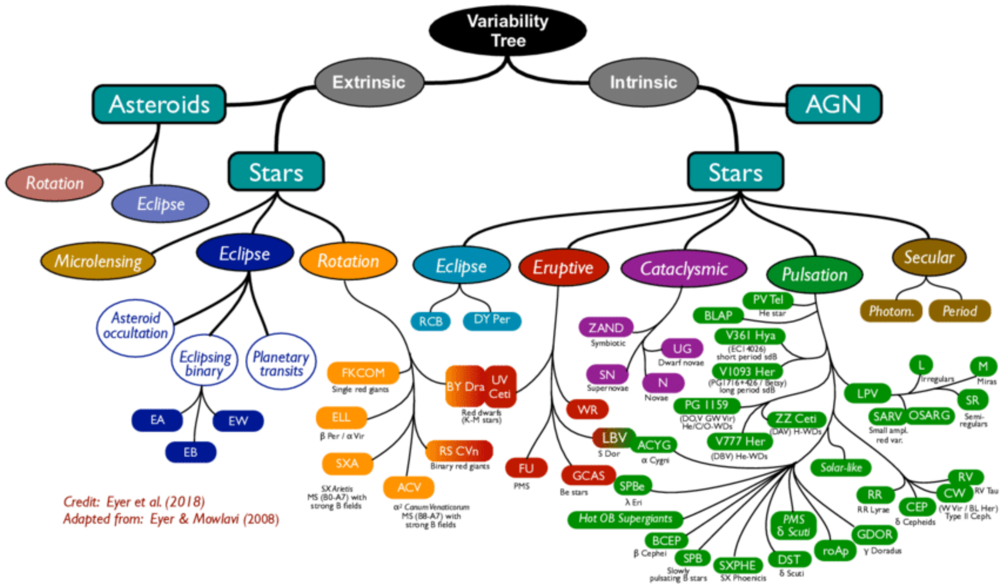
The categorisation presented in the GCVS is a foundation for understanding the diverse characteristics of variable stars. While each category has its own complex attributes and mechanisms, this introductory overview focuses on the broad classification of stellar variability. Detailed discussions and analyses of each type will be presented in subsequent sections, where the focus will shift to specific classes and the role they play in astrophysics.

Assessing a variable star often involves determining the nature of its variability; whether it is periodic, quasi-periodic, or transient. Transients, such as supernovae or micro-lensing events, exhibit photometric changes that do not repeat over observable timescales. Periodic variables are recognised by the presence of a repeating signal in their light curves, although they may also present additional irregularities in their photometric patterns.

The irregularities often hint at complex underlying physical phenomena, perhaps the interplay of both intrinsic and extrinsic factors; such as in the case of Young Stellar Objects.

1.1.0.1 Young Stellar Objects

Young Stellar Objects (YSOs) are stars in the early stages of their formation and evolution. They are primarily found in star-forming regions and are characterised by their infrared (IR) excess, strong locational dependence, and significant variability.



Credit: Eyer et al. (2018)
Adapted from: Eyer & Mowlavi (2008)

FIGURE 1.1: Showing a branching classification of variable star types (Eyer and Mowlavi, 2008)

YSOs are typically classified into different stages based on their evolutionary status:

- Class 0/I Protostars: These stars are at the earliest stages of their formation. They are heavily embedded in their natal molecular clouds. They are surrounded by dense envelopes of gas and dust and are often detected through their strong infrared and submillimeter emission.
- Class II YSOs/Classical T Tauri Stars (CTTS): These stars have shed much of their surrounding envelope and are now visible in the optical spectrum. They still possess substantial circumstellar disks from which they accrete material. This disk provides both infrared excess and many sources of intrinsic and extrinsic variability.
- Class III YSOs/Weak-lined T Tauri Stars (WTTS): These stars are more evolved YSOs that have lost most of their circumstellar disk material and show weak or no emission lines in their spectra.

The variability of Young Stellar Objects can stem from multiple sources (Wolk et al., 2018, 2013; Bhardwaj et al., 2019).

One of the most significant sources of variability in YSOs is accretion from the circumstellar disk onto the star. Accretion rates can fluctuate significantly which causes changes in the measured brightness of the star. Hot spots form on the stellar surface where material from the disk impacts, causing periodic brightness variations as the star rotates.[CITE] As YSOs rotate, spots on their surfaces (either hot spots from accretion or cooler star spots similar to sunspots) can come in and out of view, causing periodic changes in brightness. Measuring the size and relative colour of these spots through time can reveal characteristics of the spots.

The circumstellar disk itself can cause variability. Inhomogeneities cause by clumps of dust and gas within the disk can periodically obscure the star. Leading to extrinsic variability.

YSOs can often exhibit powerful outflows and jets that can interact with the surrounding material, causing variability in both the continuum and line emission observed from these stars (Guo et al., 2021). Similar to more evolved stars, YSOs can have active magnetic fields that lead to flares and other magnetic activity, contributing to their variability.

It follows that YSOs are naturally highly variable sources. These sources of variability are not mutually exclusive and so we often observe YSOs undergoing multiple sources of variability.

Despite these complexities, many periodic variables display such a dominance of periodicity in their variation that they can be effectively treated as purely periodic for practical astrophysical analysis. The study of variable stars, particularly through extensive photometric surveys, serves as a pivotal method for both understanding stellar astrophysics and mapping the structural and evolutional components of the Milky Way.

1.1.1 Intrinsic Variability

An intrinsic variable is a source whose variability stems from the internal physical mechanisms of the star. One of the most common sources of intrinsic variability is the κ-mechanism, or the Eddington valve (Eddington, 1988). This mechanism works on the principle of repeatedly storing and releasing energy within the stellar atmosphere due to changes in atmospheric opacity. The 'valve' here is thought to be governed by an inverse relationship between opacity and pressure. The abundance of He^{+} ions is thought to be the source of the kappa mechanism which is responsible for the variability found in RR Lyrae's (Gillet, 2013; Cox, 1963).

For the case of Beta Cephei variables, stellar oscillations happen in regions where temperatures approach 200,000 K, coupled with a significant presence of iron, known as the Z bump (Miglio et al., 2007).

Classical Cepheid variable stars have historically been used as the benchmark for accurate distance estimation through variability (Leavitt and Pickering, 1912; Riess et al., 2019). While these stars have been instrumental in clarifying the architecture of the Milky Way (Skowron et al., 2019; Chen et al., 2019; Matsunaga et al., 2011), their association with young stellar populations and their relative scarcity somewhat limits their utility in providing comprehensive coverage for Galactic structure tracing. Other types of variable stars also provide useful insight to mapping the Milky Way, such as RR Lyrae and Type II Cepheids. Type II Cepheids bridge the gap between classical Cepheids and the older RR Lyrae, offering insights into intermediate-age populations (Braga et al., 2020).

RR Lyrae stars are older and more uniformly distributed throughout the Galaxy, making them excellent tracers for the Galactic halo and bulge (Liu et al., 2022; Soszyński et al., 2019). RR Lyrae variables are stars on the horizontal branch of the Hertzsprung-Russell (HR) diagram, marking a phase where they burn helium in their cores. These stars exhibit variability due to pulsations, with the nature of these pulsations leading to their classification into several types. RRab stars pulsate in the fundamental mode, which affects the shape of their light curves in a specific manner (Skarka, 2014). RRc stars exhibit pulsations in the first overtone, leading to a different light curve profile (Fernley et al., 1990). There are also RRd stars, which uniquely pulsate in both the fundamental and first overtone modes simultaneously (Gruberbauer et al., 2007), and and act as excellent distance estimators (Chen et al., 2023). The mapping of RR Lyrae stars particularly allows for probing the ancient, metal-poor components of the Galaxy (Minniti et al., 2016; Savino et al., 2020). Recent studies have furthered our understanding of these stars, challenging the traditional view that associates them strictly with the Galaxy's halo. Observations have now placed some RR Lyrae stars on disk-like orbits (Matsunaga et al., 2022; Maintz and de Boer, 2005). This suggests that our current understanding of RR Lyrae star formation might be incomplete. The implication is that there might be alternative mechanisms at play, possibly involving binary star evolution, that can lead to the creation of RR Lyrae stars in such orbits (Bobrick et al., 2024). This newer viewpoint invites a broader discussion on the formation and evolution of RR Lyrae stars, suggesting that the story of these pulsating variables is more complex than previously thought. RR Lyrae variables not only deepens our knowledge of stellar physics but also enhances our understanding of Galactic dynamics and evolution. If we have a good enough indication of photometric extinction we can compare the apparent magnitude to the absolute magnitude 'M_V' calculated from equation 1.1, where 'P' is the period.

$$M_V = (-2.43 \pm 0.12) log_{10}(P-1) - (4.05 \pm 0.2) \tag{1.1}$$

Figure 1.2 shows the phase folded light curve of a W Virginis Cepheid variable.

Mira variables, with their long-period pulsations, can serve as indicators of evolved stellar populations. These objects are characterised by dramatic changes in brightness over cycles that can last up to several hundred days, these stars embody the late stages of stellar evolution for low- to intermediate-mass stars (Mattei, 1997). As asymptotic giant branch (AGB) stars, Miras exhibit pulsations that result in significant increases in their luminosity and size. These processes are driven by complex internal mechanisms, including thermal pulses and changes in molecular opacity (Joyce et al., 2024; Templeton et al., 2005). The pulsation mechanism in Mira variables is believed to be closely related to the changes in their outer layers, where varying molecular

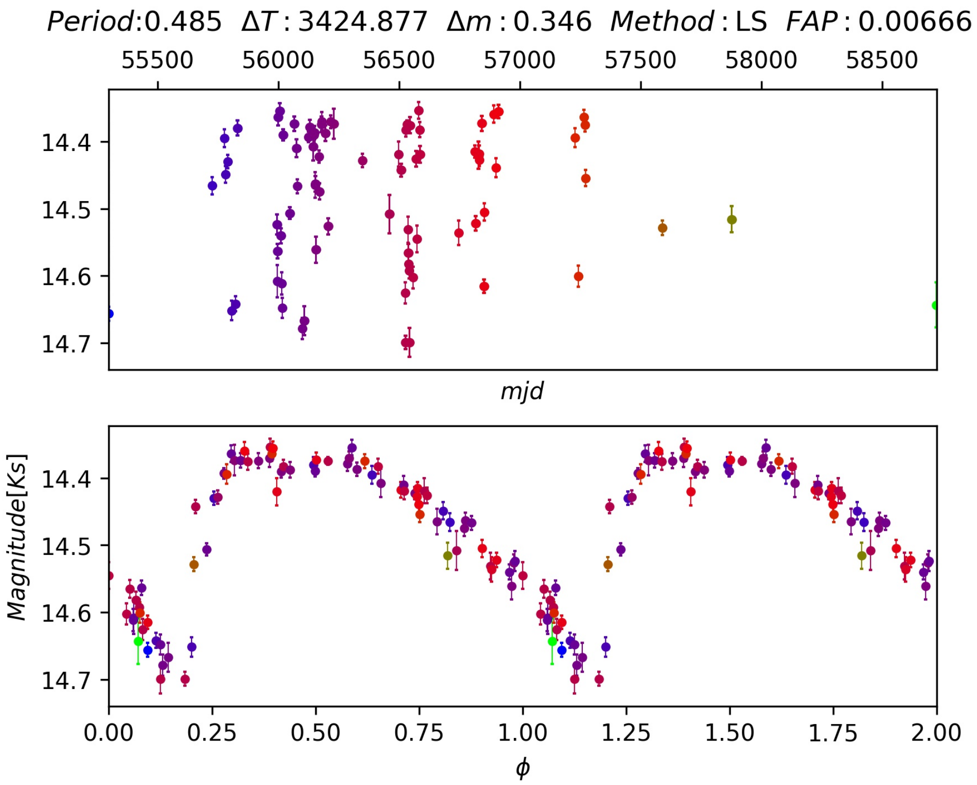


FIGURE 1.2: Showing the phase folded light curve of RR Lyrae “VVV J175038.85-174859.9”. With the coordinates 267.661° −17.817° and period of 0.583 days.

opacities lead to large-scale expansion and contraction cycles (Sanders et al., 2022). Much like many other variable stars, the inherent variability of Miras not only marks them as important objects for the study of stellar evolution but also as crucial tools for Galactic structure (Catchpole et al., 2016). Their brightness in the infrared makes them valuable tools for mapping the distribution of evolved stars across the Milky Way (Iwanek et al., 2023). The relationship between the period of pulsation and luminosity in Mira variables provides a means to estimate distances (He et al., 2021). Given their extended atmospheres and significant mass loss, Mira variables are also key to enriching the interstellar medium (ISM) with heavy elements and molecules, playing a pivotal role in the chemical evolution of galaxies (Wood, 1979; Sun Jin et al., 1982).

These are not the only sources of intrinsic variability. Stellar cold spots, which are formed the same as on our Sun, can be found on any star with an active enough atmosphere. Hot spots can also be found on stellar objects such as YSOs (Kesseli et al., 2016). These are formed via matter accretion onto the star via magnetic field lines. In both cases, the rotation of the star, and thus spot on its surface, creates a perceived stellar variability.

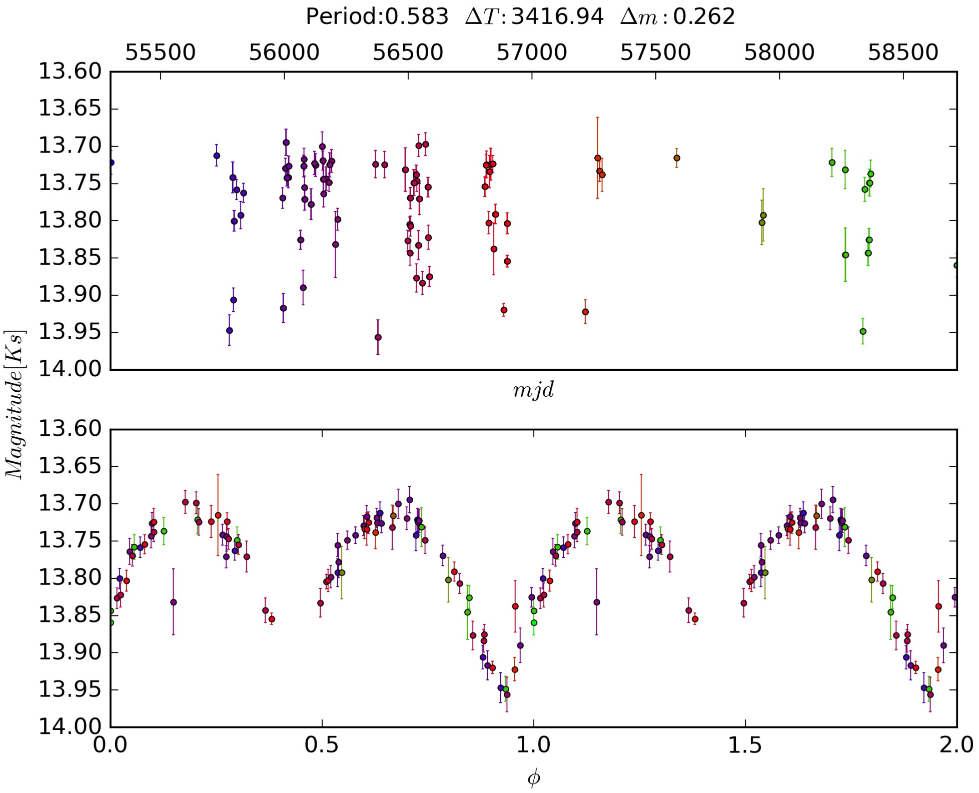


FIGURE 1.3: Showing the phase folded light curve of an Eclipsing binary of the eclipsing binary “OGLE BLG-ECL-169285”. With the coordinates 268.523° $-22.776°$ and period of 0.583 days. The period is in agreement with the Optical Gravitational Lensing Experiment (OGLE) collection of variables (Soszyński et al., 2016).

1.1.2 Extrinsic Variability

Extrinsic variability is when a source’s observed photometric variability is instead due to a third external perturber which is acting in between the observer and source. It follows that the properties of an extrinsic variable are largely dependent on their orientation with respect to the observer. For the case of eclipsing binary systems, the two stars must orbit around an axis approximately perpendicular to the line of sight of the observer. This is because their extrinsic event is the periodic obstruction of one star by the other. Figure 1.3 shows the phase folded light curve of an Algol type Eclipsing Binary found in PeRiodic Infrared Milky-way VVV Star-catalogue (PRIMVS).

Due to the nature of binary systems, simple measurable features allow for the study of multiple physical aspects of the system. Eclipsing binaries are largely independent of stellar class and so make excellent tracers for the Milky Way’s evolutionary history (Duchêne and Kraus, 2013; Chen and Guestrin, 2016).

Taken from the GCVS, eclipsing binaries are categorized into three principal types: contact (EW), semi-detached (EB), and detached (EA). In the case of contact binaries, exemplified by the well-studied W Ursae Majoris systems, both stars extend beyond their Roche lobes to share a common envelope, leading to direct contact and a state of thermal balance between the components. This forms a peanut shape. Such EW binaries demonstrate a consistent period-luminosity relationship in the infrared spectrum, a phenomenon attributed to their shared-envelope evolution. This characteristic, coupled with their minimal sensitivity to the age and metallicity of stellar populations, renders them as abundantly precise indicators for mapping the structure of the local Galactic disk and bulge regions.

Semi-detached EB-type systems, such as the prototypical Beta Lyrae, are characterized by a unique stellar interaction where only one component fills its Roche lobe. This configuration results in mass transfer from the Roche lobe-filling star to its companion, leading to dynamic evolutionary pathways distinct from those of contact binaries (Mennickent et al., 2020). The more massive recipient star accumulates matter which influences the binary's luminosity, spectral energy distribution, and period. Beta Lyrae systems are notable for their deep and complex eclipses, significant mass exchange, and the presence of accretion discs around the accretor. These characteristics make EBs crucial for understanding stellar evolution processes, mass transfer dynamics, and the role of angular momentum in close binary systems.

Detached eclipsing binaries (EA) (Eggen, 1948) are systems where both stars are well within their respective Roche lobes, avoiding direct mass exchange. This detachment ensures that each star evolves almost as if it were solitary, preserving their individual characteristics without the complexities introduced by mass transfer. Algol systems, with their distinct primary and secondary eclipses, provide a clear window into the fundamental properties of stars, including masses, radii, and temperatures.

Furthermore, the study of Algol types extends our understanding of stellar evolution, particularly in the context of mass accretion phenomena observed in some systems. This occurs when the initially more massive star evolves faster and becomes a compact object, such as a white dwarf, which can then accrete mass from its companion, leading to a reversal in the mass ratio. This aspect of Algol binaries not only contributes to our knowledge of binary star evolution but also aids in the detection of potential gravitational wave sources (Sen et al., 2023).

Planetary transits present a unique and relatively plentiful opportunity to study exoplanets through the precise measurement of time-series brightness variations of their host stars (Lissauer et al., 2023). This method hinges on the observation of slight decreases in apparent stellar luminosity which correspond to the orbiting planet passing in front of the star. The depth of this transit provides direct insight into the size of the planet relative to its host star, as the amount of dimming is proportional to the ratio of their squared radii. The repeated detection of these dimming events allows for the determination of the planet's orbital period and thus semi-major axis.

Similarly, the extrinsic variability observed in YSOs is typically attributed to the interaction between the star and its circumstellar environment. Unlike the periodic dimming caused by planetary transits, variability in YSOs can arise from inhomogeneities within their circumstellar disks. These inhomogeneities can be from dust clumps, gas accumulation, and/or disk warping, all of which will intermittently obscure the star or reflect its light, causing fluctuations in observed brightness (Lakeland and Naylor, 2022).

Such variability offers a window into the dynamic processes at play in star formation, including accretion events, disk evolution, and the early interactions between forming planetary systems and their central stars (PP2, 2023).

1.2 Time Series Analysis

Time series analysis is a critical component of studying variable stars, as it involves examining sequences of data points collected over time to identify patterns, trends, and/or periodicities. In astronomy, this analysis is complicated by irregular sampling, noise (correlated or otherwise), and other observational constraints. Techniques such as Fourier analysis, Lomb-Scargle (LS) periodograms, and machine learning methods are employed to extract meaningful insights from light curves.

In an ideal world, astronomical surveys would have strict observing sequences that are not hindered by atmospheric conditions or the lunar, diurnal and solar cycles. In this idealised case, we would have perfectly spaced observing cadences and periodic signal analysis would be simplified to standard Fourier analysis. However, this is not the case and observing cadences are highly stochastic. Robust and novel methods are required for dealing with the range of data we might expect to see.

It follows that the first efforts into period finding algorithms are Fourier based,(Deeming, 1975), followed by the least-squares approximation,(Lomb, 1976; Scargle, 1982) and more recently, its generalisation,(Vio et al., 2010) for asymmetrical light curves. Not all methods are Fourier based. Phase folding methods such as Phase Dispersion Minimisation (PDM),(Stellingwerf, 1978), string length,(Dworetsky, 1983) and the more recent conditional entropy (CE),(Graham et al., 2013b) are also prevalent (albeit not to the same degree as the Lomb-Scargle method). Phase folding methods work by first taking a list of periods with which to test. Then the light curve is folded onto itself for each period, thus producing a light curve in 'phase' space. This phase folded light curve is then analysed and some feature is extracted. This is then performed for a set of trial periods and their extracted features are compared to find a most likely period. The recent development of machine learning methods (and computers that can efficiently handle them) has also grown into time series analysis in the form of Gaussian Processes (GPs),(Angus et al., 2018).

Certain methods may be particularly useful when applied to certain stars and less useful for others. For example, the Lomb-Scargle method can be simplified as a sinusoid fitting algorithm, therefore, any star which exhibits sinusoidal-like light curves (Miras, YSOs, Cephieds...) can be analysed with ease using the Lomb-Scargle method. However, the same method applied to a star whose periodic nature is not sinusoidal (i.e. featuring 2 minima per oscillation) will likely not work. Phase folding techniques do not struggle from this issue as they are not fitting any functions. Instead, the caveat of phase folding techniques like PDM and Conditional Entropy lies with their requirement of binning a light curve in phase and magnitude space, essentially reducing the resolution of the light curve (see figure 3.5). This means that these techniques are fundamentally limited in how clearly they can identify the correct period. This also adds an extra dimension of complexity when it comes to defining the level of binning, particularly if re-analysed on a case by case basis.

There have been multiple reviews of period identification methods throughout the literature.

In (Heck et al., 1985) numerical simulations are used to compare discrete Fourier transforms, string length, and phase dispersion methods. None of these methods were found to be notably 'better'. In (Schwarzenberg-Czerny, 1999) model function and phase binning methods were compared using hypothesis-testing theory to evaluate their relative sensitivity to different kinds of signals. In this work it is shown that methods using smooth model functions, such as least squares, are more sensitive than those employing binning. Sensitivity increases for models that more closely fit features in the signal. The orthogonal multi-harmonic analysis of variance method (AOVMHW) (Schwarzenberg-Czerny, 1996) was identified as optimal. It is also mentioned that several of the methods relying on phase binning are equivalent given the same number of bins. In (Swingler, 1989) it is argued that binning methods (Phase Dispersion Minimisation) are effectively approximations to Fourier methods (i.e. Lomb-Scargle - LS). However, this makes no mention of computational limitations, which can severely impact the practical flexibility of Fourier methods.

Distefano and colleagues (Distefano et al., 2012) conducted a comparative analysis of discrete Fourier, Lomb-Scargle, and Phase Dispersion Minimization techniques to ascertain their efficacy in deducing the rotation periods of solar-like stars amidst irregular time samplings by Gaia, utilising synthetic time series. They concluded that the LS method exhibits the highest efficiency, boasting a recovery rate of approximately 60%.

It is also mentioned in (Dubath et al., 2011) that employing a singular methodology could result in a recovery efficiency of about 80%. Although, there is no mention of accuracy here. They speculated that a strategic amalgamation of these methodologies could potentially elevate this efficiency to near-perfect levels. They recommend an initial approach that involves both unweighted and weighted Lomb-Scargle techniques, the choice of which should be informed by the skewness in the magnitude distribution of the source.

Awareness of the strengths and weaknesses of these period finding methods allows for a more informed posterior analysis. In turn, the hope is that this will allow us to more appropriately adjust the methods we use to maximise robustness and minimise biases in catalogues constructed with these methods.

The Lomb-Scargle periodogram is one of the most well known period finding methods used in astronomy. The Lomb-Scargle periodogram is a Fourier-like period finding method that has been adapted to work with unevenly sampled data. A detailed discussion of the Lomb-Scargle periodogram can be found in VanderPlas (2018), which has been used as the basis for the Lomb-Scargle method in this report.

Due to its popularity, the Lomb-Scargle method is often the first method an astronomer will use when identifying a periodic signal. The Lomb-Scargle method has been shown through literature to be a semi-reliable, easy to use method (Graham et al., 2013a; VanderPlas, 2018; Swingler, 1989). However, the Lomb-Scargle method, like all other period finding methods, is not free from caveats. A common and well documented issue with the Lomb-Scargle method is its tendency to recover a multiple of the period such that $P_{LS} = N \times P_{True}$ or $P_{LS} = \frac{P_{True}}{N}$, where '$N$' is any positive integer VanderPlas (2018); Graham et al. (2013a). This problem becomes exacerbated when the Lomb-Scargle is performed over periodic data that is not analogous to a sinusoid. Periodic stars such as eclipsing binaries, especially equal mass binaries, are notoriously problematic for the LS method (Wang et al., 2012; Molnar et al., 2022).

The generalised Lomb-Scargle periodogram (GLS) has been developed as an extension of the Lomb-Scargle periodogram. The Lomb-Scargle periodogram does not include the constant term in the model (i.e. the original Lomb-Scargle method assumes an even distribution light curve about the mean).

While all phase folding methods are unique, they work in similar ways. Each of these methods only seeks to measure some quantity about a phase folded light curve. The user then finds at which period P_k gave a calculated value L_k that reflects a correct period (typically a minimum or maximum). For the case of the string length, L_k is minimised, which can be seen in equation 1.2. Where 'ϕ_k' is the phase for a given period P_k and magnitude 'm'.

$$L_k = \sum_{n-1}^{i=1} \sqrt{(m_i - m_{i-1})^2 + (\phi_{ki} - \phi_{ki-1})^2} + \sqrt{(m_1 - m_n)^2 + (\phi_{k1} - \phi_{kn} + 1)^2} \tag{1.2}$$

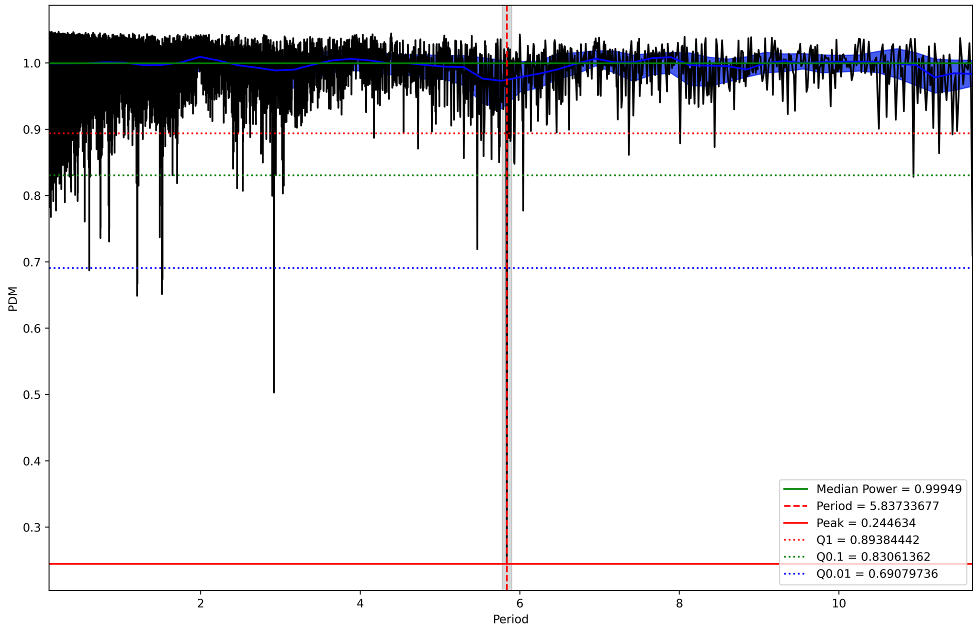


FIGURE 1.4: Showing a PDM periodogram. A dashed vertical red line has been plotted at the peak with the lowest ‘dispersion’.

Phase Dispersion Minimisation (PDM) is a phase folding technique that seeks to minimise some measure of the ‘scatter’ of data points in a light curve by phase folding the light curve at different periods. That is, the PDM method phase folds the light curve for a given list of periods and bins the phase folded light curve in both phase and magnitude space. Then it computes the variance within each bin. The period that returns the lowest total binned variance is considered to be the period recovered by this method. The Conditional Entropy (CE) method is a phase folding technique that uses a very similar method to PDM (Graham et al., 2013b). The difference being that the CE method calculates the conditional entropy for each bin.

In time series analysis, periodic signals rarely manifest as neat sinusoids, frequently adopting non-sinusoidal or Quasi-Periodic (QP) forms. Gaussian Processes (GPs) (Rasmussen and Williams, 2006) offer a sophisticated framework for modelling these intricate variations by prioritising the covariance structure within the data. The effectiveness of GPs hinges on the selection of an appropriate kernel. A kernel, or covariance function, is a function that defines the covariance between pairs of random variables within the process. Essentially, it’s a way to measure similarity or correlation between points based on their input features, which influences the shape and properties of the functions modelled by the GP. This kernel-centric methodology grants GPs the flexibility to depict the multifaceted nature of astrophysical signals. We effectively only care about relative photometry when dealing with GPs, a familiar . The ’Quasi-Periodic’ kernel effectively captures the quasi-periodic behaviour of a signal. This has shown to be effective in both the modelling of CO_2 concentrations atop the Mauna Loa volcano (Rasmussen and Williams,

2006) and the rotation periods of sun like stars(Angus et al., 2018). GPs are an example of the still evolving field of light curve analysis.

1.3 Machine Learning

Machine learning (ML) has emerged as a powerful tool in astronomy, enabling researchers to process and analyse vast datasets that were previously intractable. By leveraging algorithms that learn patterns from data, machine learning techniques are beginning to show strengths in automatically classifying variable stars, predicting stellar behaviour, and identifying novel sources/phenomena. This has opened new frontiers in astrophysics, where the complexity and volume of data continue to grow past the domain of classical analysis.

We are already seeing large changes in the zeitgeist as a product of the incredible growth in both the power and attention of machine learning. Machine learning has found footing in virtually all studies and arts (Parker et al., 2019; Lopez-Vazquez et al., 2023; Wehenkel et al., 2023; Sulc et al., 2023; Wills et al., 2023; Ramesh et al., 2021). In the field of astronomy, machine learning has already made large and unignorable changes. Figure 1.5 shows the yearly submission rate of machine learning related papers to arXiv:astro-ph. It is apparent that the rise of machine learning was a relatively fast and strong occurrence. It can also be seen that whilst machine learning was by no means popular in the 90's, it was still an active field of research in the realms of astronomy.

The concept of artificial intelligence (AI) has historical roots stretching back to at least Leibniz's *Dissertation on the Art of Combinations* in 1666. Leibniz, drawing inspiration from Descartes and Llull, proposed that a 'universal language' could represent all ideas through a limited set of fundamental concepts, suggesting that new ideas could be logically generated, potentially by a computing machine. Despite the ambitious goal ('let us calculate'), the pursuit to replicate or approximate human reasoning and the computational capabilities of the human brain continues unabated.

In 1943, McCulloch and Pitts introduced the foundational model for a computational neuron, termed the MP neuron (McCulloch and Pitts, 1943). This model is characterised by binary inputs $x_i \in \{0,1\}$ and a single binary output $y \in \{0,1\}$. An essential feature of their model is the inclusion of an 'inhibitory' input $I \in \{0,1\}$, which suppresses the output to $y = 0$ if $I = 1$. The MP neuron generates an output of y=1 or 'fires' when the summation of inputs surpasses a predefined threshold.

Despite its simplicity, the MP neuron model encapsulates a robust abstraction, enabling the computation of basic Boolean functions. Complex functions can be realised by interlinking multiple MP neurons, embodying a universal function approximator. Nonetheless, the model

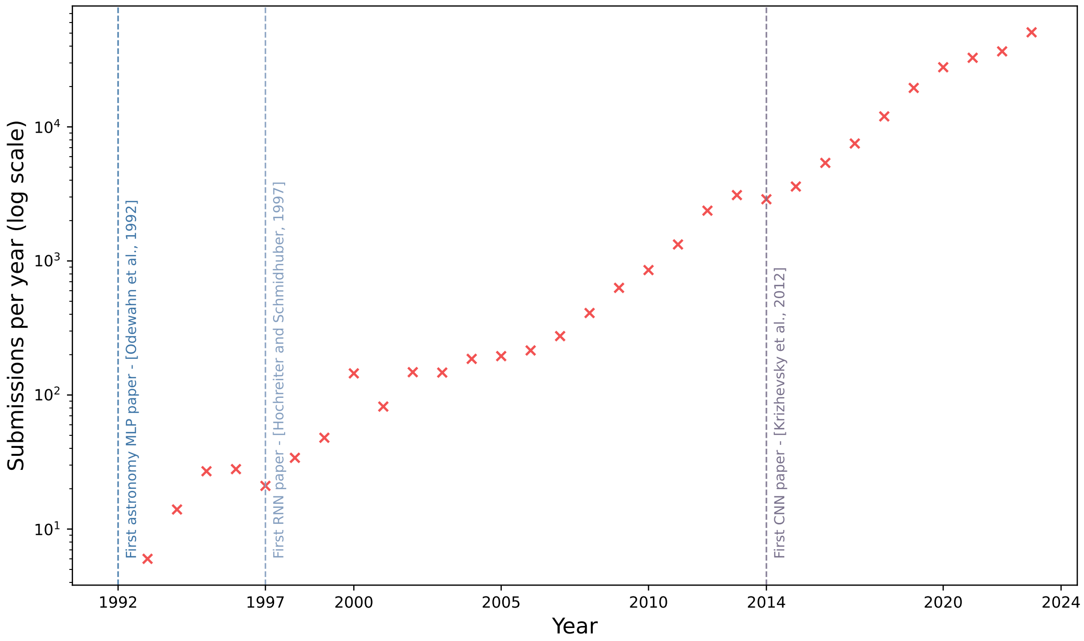


FIGURE 1.5: The yearly submission rate of arXiv:astro-ph papers related to machine learning. Significant papers are also highlighted. The full data is available here at https://www.kaggle.com/Cornell-University/arxiv

exhibits a critical limitation in its inability to learn. This gap was bridged by Rosenblatt in 1958 (Rosenblatt, 1958) through the integration of the MP neuron model with Hebb's theory of neuronal connectivity (Heb, 1950), laying the groundwork for learning in artificial neurons.

Assembling multiple artificial neurons forms a structure akin to what is depicted in figure 1.6. This configuration comprises an input layer, a 'hidden' layer acting as intermediary, and an output layer.

We can consider the task of devising a classifier capable of distinguishing between different types of stellar light curve (assuming we have an idealised set of equal length light curves). In a Multilayer Perceptron (MLP) (similar to figure 1.6), each data-point in the light curve is represented by a neuron in the input layer, with each neuron processing the data-point's numeric value and transmitting the signal onwards through the network. The subsequent neuron layer processes inputs derived from the outputs of the preceding layer. This is maintained until the signal reaches the output layer, there can be any number of 'hidden layers'. For classification tasks such, as the light curve classifier above, the output layer yields 'n' predictions, each with a value ranging from zero to one. We can then assign a set of output variable to correspond to stellar class. The desired outcome is for the network to approximate output values to their corresponding stellar class.

Standard feed-forward neural networks, such as the MLP, output a fixed-size vector for a given

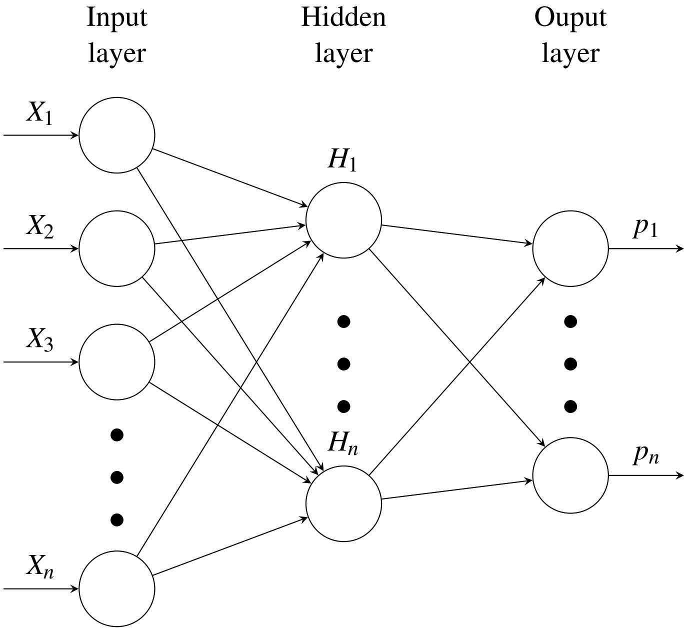


FIGURE 1.6: The MultiLayer Perceptron. The network has one hidden layer 'H1' which takes 'n' inputs of 'X_n'. The hidden layer(s) are fully connected to the input and predictive output 'p_n' which can have any number of outputs (including 1)

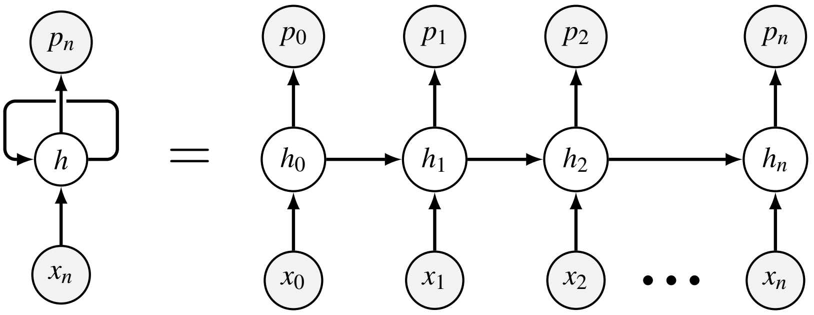


FIGURE 1.7: The Recurrent Neural Network. The unrolled Recurrent Neural Network shows how the inputs 'X_n' are passed to hidden state 'h_n to produce prediction 'p_n'

fixed-size input. However, there are scenarios where we might need to handle variably sized vectors, such as classifying a light curve dataset in the real world. A stellar light curve will vary in length due to observing patterns and data quality cuts (The VVV survey ranges from 40-1000+ measurements). This variability in length makes it challenging to use MLPs directly for classification tasks due to their requirement for fixed-size inputs. Recurrent Neural Networks (RNNs) (similar to figure 1.7) are designed to address this limitation by accepting variable-length inputs and producing variable-length outputs, thanks to their internal 'memory' that retains information from previously seen data. As an RNN processes new input, its weights are adjusted via the back propagation through time algorithm (BPTT).

We can consider the use of RNNs with our previous task of devising a classifier capable of distinguishing between different types of stellar light curve. An RNN can be employed to process the data as a sequential series $\{x_1, x_2, ..., x_N\}$, where each x_i denotes a measurement of luminosity at a given time. This series is input into the RNN in a sequentially. While the RNN outputs a prediction for each x_i, only the output corresponding to x_N—the final measurement in the sequence—is utilised for the ultimate classification p_N. The accuracy of this prediction is gauged against the actual morphological class y_N through a loss function $\mathbb{L}_N(y_N, p_N)$, which quantifies the difference between the RNN's prediction and the true classification. Optimisation of the RNN's weights is achieved by minimising $\mathbb{L}_N(y_N, p_N)$.

1.3.1 Machine Learning in Astronomy

Due in part to necessity, the overwhelming majority of machine learning research has been in the field of extra-galactic astronomy. This is largely in part to galaxies being difficult to parameterise with classical techniques. It follows that the historical path of machine learning in astrophysics is largely that of extra-galactic research. The power of MLPs were shown early with Odewahn et al. (1992). Astronomical objects were classified into stars and galaxies using data from the Palomar Sky Survey Automated Plate Scanner catalogue (Pennington et al., 1993). Their methodology involved extracting emergent image parameters such as diameter, ellipticity, area, and plate transmission from the scanned observations. These parameters served as input for training both a linear perceptron and a feed-forward MLP to segregate stars from galaxies. The most effective model demonstrated a galaxy classification completeness of 95% for objects brighter than 19.5 mag. Storrie-Lombardi et al. (1992) utilised an MLP that processed a set of thirteen galaxy summary statistics to categorise galaxies into one of five morphological types. They reported a primary accuracy of 64% and secondary accuracy reaching 90%. Early on it was shown that stellar spectra can be classified by temperature (von Hippel et al., 1994) and spectral type (Klusch and Napiwotzki, 1993). These initial investigations further affirmed the capability of MLPs as useful tools in the automatic classification of astronomical objects.

By 2014, Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012) were being applied to the search for pulsars as part of a broader ensemble of methods (Zhu et al., 2014). The effectiveness of their approach is shown by achieving virtually perfect results. Where 100% of pulsar candidates from their test set were ranked within the top 961 out of 90,008 candidates. The classification of spectra into quasars, stars and galaxies was shown to be an effective application of a 1D-CNN (Hála, 2014). A notable win for the power of machine learning was found in leveraging the Galaxy Zoo dataset (Raddick et al., 2010) where it was possible to construct a training set from citizen labelled data to classify galaxies. This was shown in Dieleman et al. (2015) and later Walmsley et al. (2020).

The advent of Recurrent Neural Networks (RNN, Hochreiter and Schmidhuber, 1997) in astronomy marked a significant milestone, with the first known application being the classification of potential extraterrestrial narrow-band radio signals (Brodrick et al., 2004). Many notable works have contributed to underscoring the efficacy of RNNs in astronomical classification tasks (Naul et al., 2018; Gomez Gonzalez et al., 2018; Carrasco-Davis et al., 2019; Finke et al., 2021).

1.3.2 Machine Learning for the treatment of light curves

In the Supernova Photometric Classification Challenge (SPCC), Karpenka et al. (2013) presented a neural network approach that relied on fitted light curve coefficients, proving competitive against more complex models. Their method fits supernova light curves across various filters to a highly adaptable analytical function, with the aim to capture the characteristics of supernovae types. This fitting process, coupled with a comprehensive feature vector that includes parameters such as the number of flux measurements and the maximum likelihood of fit, enables the neural network to differentiate between Type Ia and non-Ia supernovae with 82% accuracy.

The nature of RNNs mean they are often adopted when unevenly sized sequential data is used - light curves.

RNNs demonstrated the ability to accurately inpaint missing time series data from the NASA Kepler mission (Capizzi et al., 2012). The noise profiles and covariances of stellar light curves are complex and simplified for human interoperability when simulated (Emmanoulopoulos et al., 2013). Thus it is notable that the RNN here could *understand* the data so well.

This interest in RNNs for regression tasks grew notably in the late 2010s, with (Shen et al., 2019; Morningstar et al., 2019; Liu et al., 2019) showcasing their utility in denoising gravitational wave data, reconstructing gravitationally lensed galaxies, and predicting solar flare activity, respectively. These studies collectively reinforce the role of RNNs as effective tools for regression in astronomical time series analysis.

The use of RNNs for auto-encoding has also shown to be possible for the embedding of variable main sequence stars (Kügler et al., 2016). An echo state network (Jaeger, 2002), was used to extract representation embedding of variable main sequence stars, uncovering emergent properties such as temperature and surface gravity.

With the advent of increasingly comprehensive astronomical surveys, the necessity for automated systems capable of efficiently extracting and classifying variable objects has surged. This requirement aligns well with the capabilities of machine-learning classification algorithms, where photometric light curves serve as the basis for feature generation, and training data are sourced from established variable star catalogues. The GCVS exemplifies such a catalogue, having compiled variable objects since 1946. This approach has been applied to data from All-Sky Automated Survey for Supernovae (ASAS-SN) (Jayasinghe et al., 2018, 2019), Gaia

DR2 (Rimoldini et al., 2019), EROS-II light curves (Kim et al., 2014), and in the identification of microlensing events (Husseiniova et al., 2021).

Supervised learning methods demand extensive, pre-classified datasets that encompass the diversity of objects a new survey may observe. Training set biases are regularly reflected in the classifier's performance. This often results in the compound effects of biases and in many ways, the result of such a self supervised tasks. We end up with inheriting the biases of both.

For example, using a Gaia light curve training set (such as Rimoldini et al. (2023)) to classify a WISE light curve dataset would be problematic. We would essentially inherit the optical depth of GAIA with pixel scale and light curve quality of WISE, essentially resulting in the worst features of both. This issue is not critical, if we are able to parameterise the data using the biased limitations of the training data (i.e. EB stars will likely be sufficiently parameterised in such an example). It follows that a robust understanding of the data used for self supervised machine learning is of paramount importance.

In Dubath et al. (2011) a random forest method is used to automatically classify periodic variables in the 'Hipparcos' data. In this paper, it is noted that the most important feature for correct classification was the period. In Kim and Bailer-Jones (2016) a package for the automatic classification of periodic variables is shown. Much like the previous paper, the period is by far the most important feature when classifying such stars. It follows that the performance of these machine-learning methods are directly affected by the accuracy and reliability of their period finding methods. However, not all machine-learning techniques require a calculated period, some works strive to classify without computing features. In Becker et al. (2020) a featureless classification method is used.

Unsupervised learning algorithms, like clustering, adopt a more neutral stance, capable of uncovering dataset structures and identifying anomalies (Brett et al., 2004; Eyer and Blake, 2005; Sarro et al., 2009). More nuanced self supervised techniques have also shown to be incredibly powerful and are primed to fundamentally change astronomy (Hayat et al., 2021; Sarmiento et al., 2021; Slijepcevic et al., 2022) and more generally, science (Smith et al., 2023; Bountos et al., 2022; Saleh et al., 2022; Lastilla et al., 2022) Considering the size and complexity of future datasets, it seems compelling that unsupervised learning will play a very prominent role in processing and analysis. Employing a blend of supervised and unsupervised learning may offer the best strategy for managing and analysing burgeoning datasets.

1.4 The Vista Variables in Via Lactea (VVV) survey

The United Kingdom Infrared Telescope (UKIRT) Infrared Deep Sky Survey (UKIDSS) (Lawrence et al., 2007) is a large-scale near-IR survey conducted on the United Kingdom Infrared Telescope (UKIRT). UKIDSS, which began operations in May 2005, serves as a successor to the Two Micron All Sky Survey (2MASS) and surveyed 7500 $\deg^2$ of the Northern sky. UKIDSS consists of five individual surveys, of which, the UKIDSS Galactic Plane Survey (GPS) is one. GPS surveyed 1868 $\deg^2$ of the Galactic plane with Galactic latitudes $|b| > 5^{\circ}$ in the J, H and K filters. The GPS provides two/three epochs of K-band photometry and has an aim of investigating phases of stellar evolution via the detection of high amplitude near-IR variability (Contreras Peña et al., 2014). The GPS will investigate eruptive young stellar objects (YSOs) also known as FUor and EXor events (Montmerle, 1990; Contreras Peña et al., 2014; Lucas et al., 2017). As GPS is a large area with only two epochs, it is suitable for the study of such high amplitude variability with long decay times.

A main caveat of GPS is the limited epochs. Hence, differentiating between FUor events and other sources of variability, such as Miras and Novae, can be difficult. Another caveat of GPS is the relatively low dynamic range. The conservative saturation limit for GPS is $m_k < 12.0$, $m_H < 12.75$ and $m_J < 13.25$ and a 90 % completeness of $m_K = 18.0$, $m_H = 18.75$ $m_J = 19.5$ (Lucas et al., 2008). If we consider that an average FUor event has an increase in brightness of $\approx$ 6 mag (Bell et al., 1995) this significantly reduces the range of FUor events that can be fully studied. Furthermore, events with a shorter decay time are at risk of not being detected, such as EXor events. In Contreras Peña et al. (2014) it is discussed that GPS was largely used as a precursor for the VISTA Variables in the Vía Láctea (VVV) survey (Minniti et al., 2010).

The VVV survey is a infrared time-series survey focused on the southern viewable Galactic disk and bulge. The VVV survey feature light curves with approximately 40 to 80 epochs of data in the K_S band (from 2010 to 2015). In addition to the K_s band data VVV also features 2 epochs of $Z, Y, J \& H$ photometry taken at the start and end of the survey. The VVV survey saturates at $K_S \approx 12$ and is 90 % complete at $K_S \approx 16.8$ mag however it is noted the magnitude limit is strongly dependent on the crowding of the field. Due to its multiple epochs in and around the Galactic centre, VVV is useful for a wide array of studies, particularly in time-domain astrophysics.

The VISTA telescope features an array of 16 Raytheon VIRGO HgCdTe 0.84-2.5 micron detectors (Bornfreund, 2005) in an arrangement shown in figure 1.8. For a standard K_S 'tile' measurement VISTA will start by taking 2 4-second 'jitter' images shifted by < 30", this is performed to account for the $\approx 2\%$ 'bad' pixels for a typical VISTA detector.

VISTA moves the telescope to 6 unique pointing positions to fill in the gaps between the detectors. The telescope is moved in 3x2 pattern with 3 positions in the Y-axis of the camera and 2 in

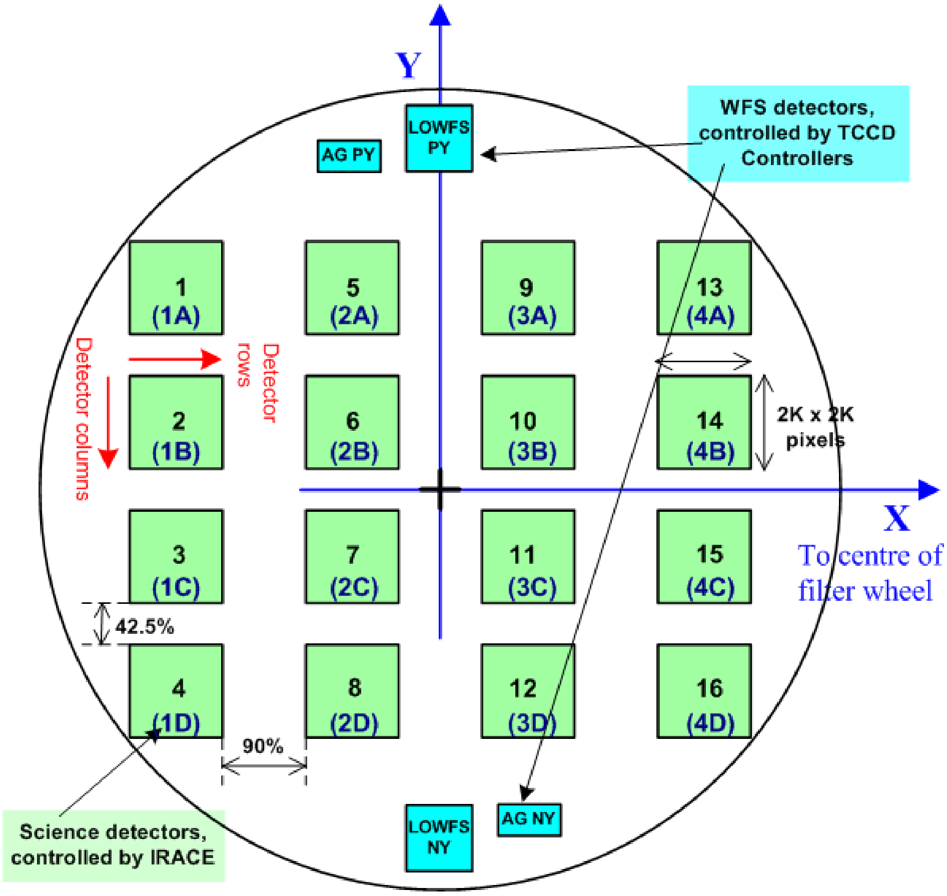


FIGURE 1.8: Diagram representing the detector arrangement as if looking directly down the camera body. The +Y-axis here points north and the Z-axis would point towards the sky

the X-axis. Each of which is a stacked pair of jitter images. This method creates full coverage of a 1.5 deg coverage of the night sky. This also means that, due to overlap, multiple sources will be measured multiple times for a single observing tile, for this report, these will be referred to as ‘paw-print pairs’. Figure 1.9 shows the distribution of measurements in a single VISTA tile.

The VVV survey was extended into the VVV eXtended survey (VVVX). VVVX furthers the original survey in both the original area and expanded regions. With an additional 3 to 10 K_s observations in the original VVV area. VVVX covers 1700 deg^2 and will take 2000 observing hours. Figure 1.10 [1] shows the original VVV area (red) and its extended regions. The blue region extends into the Vista Hemisphere Survey (VHS) (McMahon et al., 2013).

VIRAC is a near-infrared astrometric catalogue of the VVV survey. VIRAC features light curves for over 80 million unique sources over a 560deg^2 area. For this project, the VIRAC2 catalogue (Smith et al. in prep) is used which differs from that discussed in Smith et al. (2018) in the way its photometric reduction was performed. The original VIRAC was based on the standard

[1] vvvsurvey.org

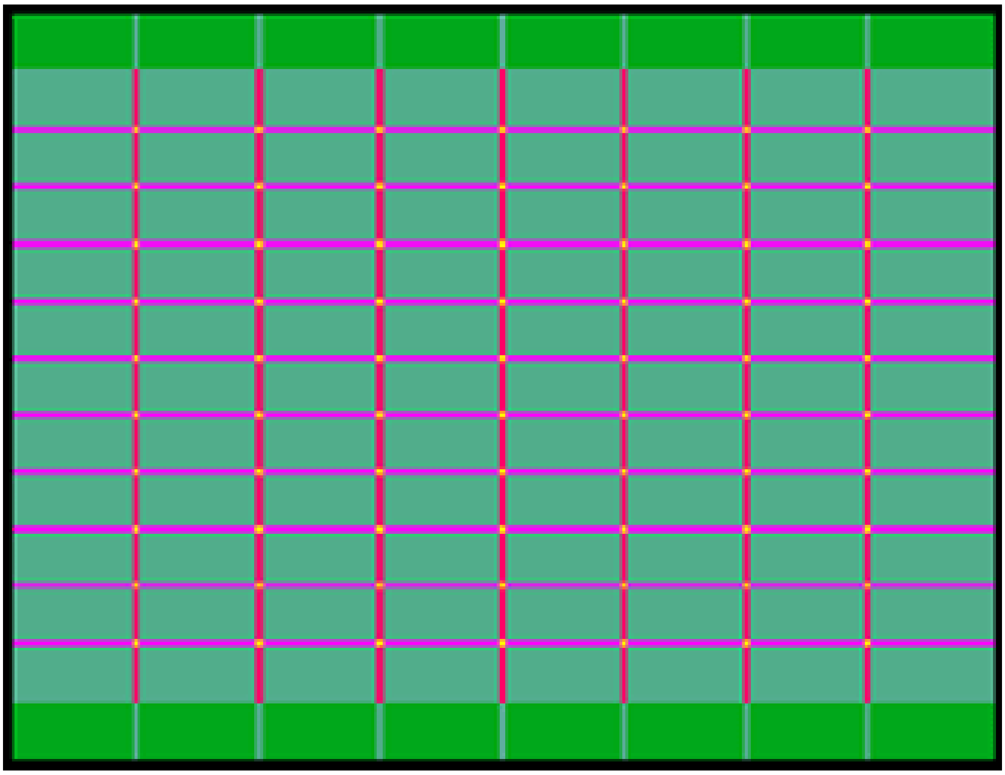


FIGURE 1.9: Diagram representing the number of detections within a VISTA tile. **dark green** = 1, **light green** = 2, **magenta** = 3, **red** = 4, **yellow** = 6, in the unit of a pair of jitter images.

products provided by the v1.3 pipeline of the Cambridge Astronomical Survey Unit (CASU) while VIRAC2 uses profile fitting photometry carried out with DoPHOT (Schechter et al., 1993). The new procedure was implemented to mitigate the blending of sources in highly crowded regions of the inner bulge (Hajdu et al., 2020).

As VVV is only near-IR, classifying the sources based on colour alone is sometimes not sufficient. It is stated in Contreras Peña et al. (2017) that AGB stars are the largest source of contamination when studying YSOs using VVV. Much like in UKIDSS, differentiating between YSOs and AGB stars can be difficult (Guo et al., 2020; Koenig and Leisawitz, 2014). Both YSOs and AGB stars have IR excess caused by surrounding material being heated by the star. In the case of YSOs, this material is the circumstellar disk. AGB stars feature a thermal pulse cycle where they radially expand and contract. This creates fluctuations in effective temperature and subsequently, optical opacity.

AGB stars radiate in IR due to the thick circumstellar envelopes. However, when we consider the extra dimension of time afforded to us by multiple epochs of data we can drastically reduce the degeneracy of such stellar classifications. In the case of differentiating between YSOs and AGB stars, AGB stars typically have longer periods compared to YSOs. AGBs periods are typically around 100-1000 days (Takeuti et al., 2013; Chibueze et al., 2016) with the more optically obscured OH/IR stars having periods between 500-1800 days (Jiménez-Esteban et al.,

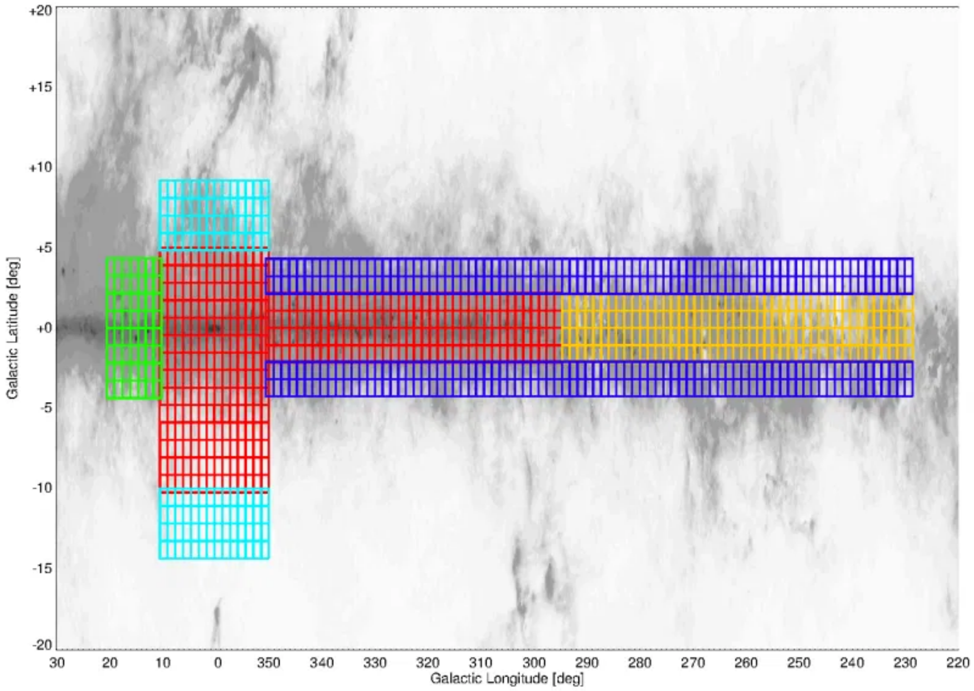


FIGURE 1.10: Showing the a colour coded map of the area covered by VVV and VVVX. **Red**: The original VVV area. **Cyan**: The extension to the original bulge area. **Cyan**: The extension to the original disk area by 10 degrees Galactic longitude ($b = \pm4.5$ deg). **Yellow**: The extension to the disk area from $l = 295$ deg to $l = 230$ deg ($b = \pm2.28$ deg). **Blue**: Extending the southern disk area laterally by 2.22 deg. In doing so, partly overlapping with VHS.

2006) while YSOs typically have periods between 0.2-14 days (Wolk et al., 2018),Although this is dependent on the source of variability, such as AA Tau like extinction events (Covey et al., 2021).

Furthermore, AGB stars undergo other perturbations which effect their periodicity on long time scales (Höfner and Olofsson, 2018) and so light curves with sufficient temporal coverage (such as those found in VVV) will look less perfectly variable when compared with a YSO. Time series analysis can be performed both manually via visual inspection of a light curve or automatically by the extraction of features such as the period and amplitude (or less obvious features such as the 'Q' and 'M' values (Cody et al., 2014)).

Chapter 2

The verification of periodicity with the use of recurrent neural networks

Abstract

The ability to automatically and robustly self-verify periodicity present in time-series astronomical data is becoming more important as data sets rapidly increase in size. The age of large astronomical surveys has rendered manual inspection of time-series data less practical. Previous efforts in generating a false alarm probability to verify the periodicity of stars have been aimed towards the analysis of a constructed periodogram. However, these methods feature correlations with features that do not pertain to periodicity, such as light curve shape, slow trends and stochastic variability. The common assumption that photometric errors are Gaussian and well determined is also a limitation of analytic methods. We present a novel machine learning based technique which directly analyses the phase folded light curve for its false alarm probability. We show that the results of this method are largely insensitive to the shape of the light curve, and we establish minimum values for the number of data points and the amplitude to noise ratio.

2.1 Introduction

The identification of periodic variable stars is not a trivial task; well-understood statistical measures can be used to identify variability in time-series but not so easily periodic variability. The Stetson variability index 'I' (Stetson, 1996) compares the variability of each observation with its neighbour and their errors. The Von Neumann eta index 'η' (Neumann, 1941) represents the ratio of the mean of the successive differences squared, to the variance of the light curve. Both of these methods are reasonably robust in detecting variability in time-series. More simplistic methods, such as a comparison between some measure of scatter (Interquartile Range, Standard deviation σ or Median Absolute Deviation) and the uncertainty, have also been shown to be useful (Sokolovsky et al., 2017). Using tools such as the Lomb-Scargle method (Lomb, 1976; Scargle, 1982) and Phase Dispersion Minimisation (PDM, Stellingwerf, 1978), we can construct a periodogram to probe for periodic variability. Nevertheless, extrema in the periodogram are likely to be present regardless of whether or not the source is truly periodic. These extrema can scale with the amplitude of the periodic signal such that periodograms of periodic sources become distinct from truly random variability. However, in cases where a light curve features aperiodic or secular variability, ambiguities can arise (Park et al., 2021). This is of particular issue when dealing with stars which can feature multiple sources of variability, such as asymptotic giant branch stars (Templeton et al., 2005), whose long term periodicity could be undifferentiable to that of secular variability by periodogram analysis alone. Furthermore, their values do not scale universally (i.e. the peak value for an aperiodic source may be the same as that for a periodic source).

In cases where extrema are not present, this could be interpreted as an indication of insufficient periodogram coverage or the lack of periodic variability.

Thus, we do not automatically obtain a universal measure of periodicity from a periodogram. If a periodogram shows candidate periods, then for smaller selections of sources, it is feasible to manually verify the periodicity of each. This is typically performed by visual inspection of the phase folded light curve. Looking forward, in the current and future age of survey astronomy with surveys such as LSST (Ivezić et al., 2019), ZTF (Bellm et al., 2019), Kepler (Borucki et al., 2003) and TESS (Ricker et al., 2015), we anticipate time-series catalogues of sizes that render sufficient manual inspection increasingly non-viable. Hence, a reliable and robust metric for identifying periodicity is required.

It is not a guarantee that a large survey will feature high cadence sampling. Surveys such as VISTA Variables in the Via Lactea (VVV, Minniti et al., 2010; Saito et al., 2012), the NEOWISE mission of the Wide Field Infrared Survey Explorer (Wright et al., 2010; Mainzer et al., 2014) and *Gaia* (Gaia Collaboration et al., 2021) have catalogues which can also feature large sample sizes for which rigorous human inspection is impractical. These surveys contain relatively few observations for each source, an issue that is also very common in small, targeted observing projects. The sparse sampling makes it harder to confirm periodicity with classical methods.

The metric for determining periodicity in a time-series is commonly referred to as a False Alarm Probability (FAP). Previous work on determining an accurate FAP has largely been directed toward the analysis of the constructed periodogram. These methods, such as the method proposed by Baluev (2008), employ extreme value statistics to determine an upper bound for the false alarm probability of a Lomb-Scargle periodogram. This has the clear limitation that the method is designed to distinguish sinusoidal variations from Gaussian white noise, not accounting for stochastic variability, non-Gaussian errors, imprecise error estimates and non-sinusoidal periodic variations. Baluev (2009) extended their earlier work to the case of multi-harmonic light curves but this is only a partial solution to the above issues. Bootstrapping is another commonly used technique where the periodogram of a light curve that has been randomly shuffled N times to create N aperiodic periodograms is compared to that of the unshuffled light curve. The FAP in this case is the percentage of times the peak of an aperiodic periodogram is larger than that of the peak from the suspect periodic periodogram.

In Stellingwerf (1978), a statistical analysis of the constructed PDM periodogram is used to obtain a metric of false alarm probability (P-value). This method assumes that photometric errors are perfectly estimated Gaussians. The absence of any other aperiodic variability is also assumed. There is also no treatment of spurious artificial periodic signals, which can occur with unevenly and sparsely sampled light curves. Many surveys feature these periods at varying rates of incidence. It is of particular note for ground-based surveys with semi-regular observing patterns, such as the VVV survey. Methods such as PDM that bin the phase-folded light curve to construct their periodogram are also limited by imperfections in the model. This can become increasingly significant as sampling decreases. This issue exists even with the binless approach to PDM presented by Plavchan et al. (2008). Separately, heuristic methods based on reduced χ^2 statistics have been employed to distinguish true and false periodic variable star candidates (e.g. Irwin et al., 2009). This explicitly acknowledges the effects of an imperfect light curve model and imprecise photometric uncertainties. In this work, we show how we can utilise neural networks to differentiate between true and false periodic variable star candidates without the need for a prior light curve model.

2.2 Method

In our approach, the analysis of the light curve is achieved via a Recurrent Neural Network. An RNN was chosen because they are designed and used for serially correlated data, such as astronomical light curves. Previous efforts in their use with light curves have shown their applicability and ability to parse astronomical time-series data (Burhanudin et al., 2021; Zhang and Zou, 2018). This network is trained on pre-labelled periodic and aperiodic phase-folded light curves of variable stars. The network was trained for 96 epochs[1] with an Adam optimiser (Kingma and Ba, 2014) and with 20% of the training data used as a validation set. Early stopping was used to halt training as soon as the incremental change in the validation loss function, $\Delta L < 10^{-5}$.

The model is constructed with 13 Gated Recurrent Unit (GRU, Cho et al., 2014) layers, 1024 nodes per layer and a binary cross entropy loss model. The choice of GRUs over Long-Short Term Memory (LSTM, Hochreiter and Schmidhuber, 1997) was motivated by the calculated loss, which was lower for GRUs. The RNN was written in Keras (Chollet et al., 2015).

Ablative testing has shown that the specifics of the architecture of the network are not crucial and having ‘enough’ GRUs is sufficient for operability.

2.2.1 Data preparation

The input training data consists of the magnitude (m_i), phase (ϕ_i) and change in phase (i.e. $\Delta\phi_i = \phi_i - \phi_{i-1}$). Magnitude errors were not used for this method as they commonly do not fully represent the true photometric uncertainty. We tested various combinations of features and removing the magnitude error consistently improved performance with lower loss and higher accuracy. Instead, we reject any points with a large magnitude error ($m_{i,err} \geq 0.1$ in this case). We also reject points with a high DoPHOT (Schechter et al., 1993) ‘Chi’ parameter, which indicates a poor fit to a stellar profile.

The input also includes a feature that is derived from an interpolated fit of the time-series with 200 evenly spaced points, performed by an inverse distance-weighted K-nearest neighbours (KNN) regressor (Fix and Hodges, 1951) which was taken from Scikit-learn (Pedregosa et al., 2011). This was performed as a form of smoothing in an attempt to more clearly display variability with evenly spaced data.

A randomly variable light curve will have an interpolated fit that tends towards a straight line. Each of these features were added after ablative testing (i.e. features were added and removed iteratively and the combination of features that produced the highest accuracy and lowest loss was used). Each light curve was either cut to 200 data points in size or padded with zeroes to a length of 200.

The same light curve is phase shifted randomly 10 times by an amount between $0\,\text{and}\,2\pi$ and each version is shown to the neural network. This is done in an attempt to remove a dependency on the starting position of the light curve. This is similar to the methodology for contrastive learning (Chen et al., 2020). We do not want the network to care about the absolute phase value.

[1] An ‘epoch’ here is an iteration over the whole training set

Alternatively, we could ensure the light curve is always ordered from a set point in the light curve, such as the turning points. However, we found this step to be unreliable with noisy data. A single unfiltered outlier or otherwise erroneously extreme point would cause such an approach to fail as the light curve's minima could be incorrectly identified.

2.3 Data

The training data used for training the neural network FAP (NN FAP) is a combination of both real and synthetic light curves.

In the trained model used for this paper, there were 20 000 real and 60 000 synthetic light curves with half of each corresponding to periodic or aperiodic. This means that a FAP of 0 was given to the 10 000 real and 30 000 synthetic periodic light curves and a FAP of 1 was given to the other 10 000 real and 30 000 synthetic aperiodic light curves. The synthetic light curves were split evenly across each of the five listed equations (2.1–2.4).

Through the development of this method, it was found that a small number of mislabelled light curves can have a large impact on the abilities of this method (i.e. an aperiodic light curve being labelled as periodic or vice versa).

2.3.1 Real training data

The training data are VVV light curves whose periodic nature was supported by classification from two optical surveys. A set of 10,000 known real periodic light curves were identified by eye (by co-authors NM, CM & WC) after cross-matching data from the VVV survey, (and a pre-release version of its time-series catalogue, VIRAC 2-β (Smith et al., 2018, Smith et al., in prep) with other known periodic variable star catalogues, namely the ZTF catalogue of periodic variable stars and the ASAS-SN catalogue of variable stars (Chen et al., 2020; Pawlak et al., 2019). The cross-matching was performed to generate a list of suspect periodic and aperiodic variable stars.

All of the 10,000 aperiodic light curves were identified by eye as rejected periodic variables.

Figure 2.1 shows a random selection of real training light curves and their interpolated fit. Both the interpolated fit and the raw magnitude measurements are given to the RNN.

Figure 2.2 shows the distributions of number of data points, signal-to-noise ratio and number of cycles in the time series for the training data. The real data is drawn directly from this distribution.

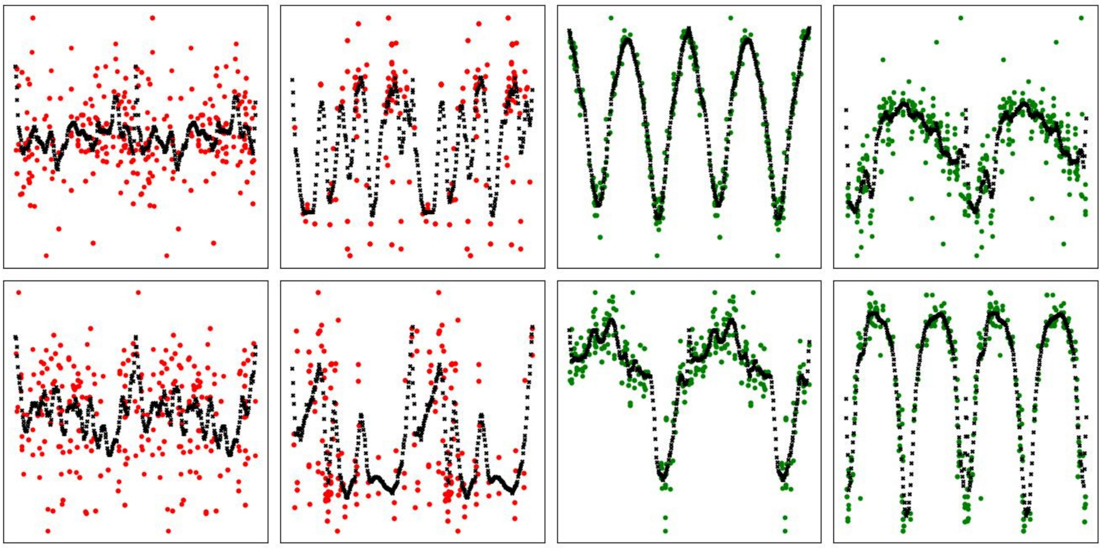


FIGURE 2.1: Some example real aperiodic (red) and periodic (green) light curves used for training. The black points represent the evenly spaced fit provided by the KNN regressor.

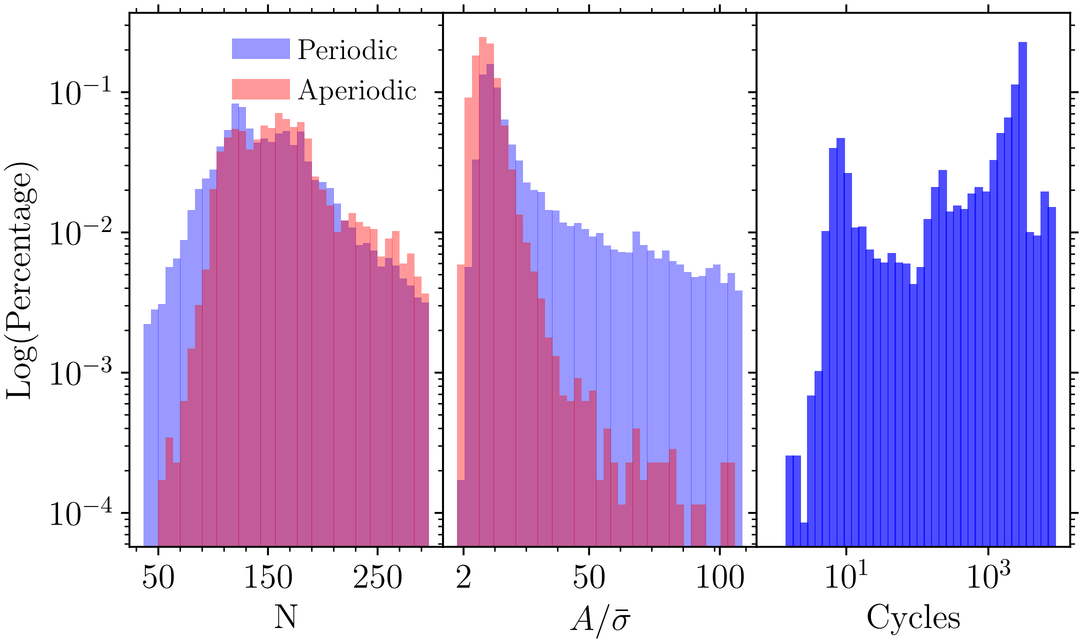


FIGURE 2.2: Showing the distribution of the number of data points (N), signal-to-noise ratio and cycles (light curve length / period) for the training set used.

2.3.2 Synthetic training data: Periodic light curves

Synthetic light curves were created via the use of a real light curve with a periodic signal injected, similar to the work by Graham et al. (2013a). An overview of the steps taken are as follows:

1. Remove all photometric information from a real light curve, retaining only the time stamps.
2. Inject periodic signal into 'blank' light curve **(see equations 2.1–2.4)**.
3. Generate the errors by sampling from those associated with the real photometry, using a look-up table.
4. Scatter light curve based on the injected error.

Training the neural network on exclusively sinusoidal light curves could bias our FAP against Eclipsing Binaries and other more complex light curves.

Figure 2.3 shows an example of each of the forms of light curves generated with 1000 measurements with an amplitude of 1 mag. These equations aim to roughly (but not exactly or comprehensively) model the common types of pulsators and binary light curves that are seen (Molnar et al., 2022). Type 1 is a distorted sinusoid which is a fairly standard form for synthetic light curves (Cincotta et al., 1995; Huijse et al., 2012). Types 2 & 5 are eclipsing binary-like light curves (i.e. more than one turning point per period). Type 3 is used to mimic the common identifying feature of a contact binary system (Kirk et al., 2016) and Type 4 is a simple sinusoid. An important reason for using multiple shapes to train the network is to remove as much of a dependency on light curve shape as possible. This is similar to the methodology for contrastive learning. By showing the network multiple different shapes of a periodic signal we aim to remove any biases related to its shape.

The method by which the synthetic data is created also means that the light curve parameters are drawn from the distributions shown in Figure 2.2. The periods used are randomly selected from a uniform distribution between 0.1 and half the length of the light curve (period $\sim\mathbb{U}(0.1, \sim1500)$). We note that the period (or number of cycles) used for the synthetic light curves is largely inconsequential to how it is perceived by the RNN. The RNN is only shown the phase fold of the light curve and so there is little difference between otherwise identical light curves with different periods. This is also the case for the total time range of the light curve. Provided at least one cycle is captured, the number of measurements and signal-to-noise are the limiting factors. This is a potential caveat for this method as a low FAP could be assigned for a light curve with only one cycle, which is not sufficient for the actual identification of periodicity. We

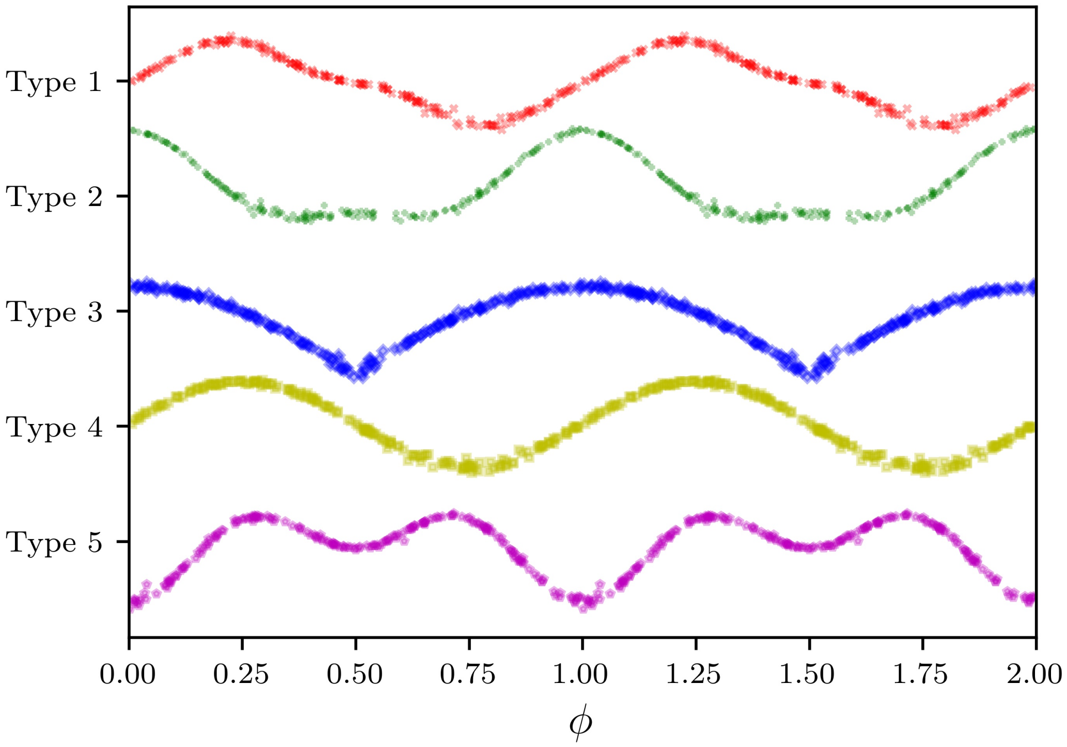


FIGURE 2.3: Examples of each of the forms of the light curves used for testing and training the neural network FAP.

recommend only trusting the FAP from this method if the period is less then half of the length of the light curve (i.e. at least two cycles are captured).

$$\texttt{Type 1.}\ m(t) = 0.5sin\left(\frac{2\pi t}{P}\right) - B_1 sin\left(\frac{4\pi t}{P}\right) - B_2 sin\left(\frac{6\pi t}{P}\right) \tag{2.1}$$

$$\texttt{Types 2 \& 5.}\ m(t) = 1 \pm \left(A_1 sin\left(\frac{2\pi t}{P}\right)^2 + A_2 sin\left(\frac{\pi t}{P}\right)^2\right) \tag{2.2}$$

$$\texttt{Type 3.}\ m(t) = \left|sin\left(\frac{2\pi t}{P}\right)\right| \tag{2.3}$$

$$\texttt{Type 4.}\ m(t) = sin\left(\frac{2\pi t}{P}\right) \tag{2.4}$$

A periodic signal is added to the source light curve, and the photometric error is derived using a KNN search of a dataset containing information about the photometric uncertainty of 1 000 000 data points from the VIRAC database. This dataset is utilised to identify the 100 nearest neighbours, from which the mean and standard deviation are computed.

Each data point in the light curve has its photometry (m) and photometric error (m_{err}) drawn from a Gaussian constructed of these 100 nearest neighbours.

2.3.3 Synthetic training data: Aperiodic light curves

We employ two methods to generate aperiodic light curves: a real or synthetic periodic variable has its photometric order randomly shuffled. The time data is left unmodified to conserve the observing cadence of the original survey. We effectively create a light curve of random noise with the survey's observing pattern conserved. This method also removes any other correlated effects, such as photometric uncertainty, that may be present in real aperiodic light curves. One caveat present is that by destroying correlated effects, the neural network could differentiate between the aperiodic and periodic synthetic light curves with greater ease. The second method of aperiodic synthetic light curve generation involves taking a known non-variable star (identified with a Stetson index <0.1) and re-sampling the photometric points with a larger scatter. For each measurement a Gaussian is constructed with $\mu = m_i$ and $\sigma \geq 3 \times m_{i,err}$, where $m_{i,err}$ is the measurement error. The light curve is then re-scaled to ensure a realistic amplitude.

This method retains as much temporally correlated, but non-periodic, information as possible compared to the random shuffle method. An example of this is with astronomical seeing, which can vary on long timescales, affecting multiple measurements. With VVV (and subsequent catalogue VIRAC 2-β) data we have instances where bad seeing causes DoPHOT to systematically underestimate flux in crowded fields. Such a case could appear as a non-periodic signal in the light curve. In less crowded fields, poor weather will increase the uncertainty at times, creating correlated uncertainty which may occasionally lead to a spurious aperiodic signal. This is of particular note as the neural network is never shown the photometric uncertainty. This method of inflating measurement error will weaken but not fully destroy these correlated effects.

The random shuffle method enables training with non-Gaussian aperiodic signals. Due to the limitations of these methods, it is beneficial to also have real training data. The synthetic data has the advantage of volume with the certainty of aperiodicity. This allows us to construct a training data set large enough to train an RNN.

2.3.4 Test Data

We generate 3 data sets to test our classifier. A real data set was constructed by manually classifying 8000 previously unseen real light curves taken from the same VVV survey. These 8000 sources were identified from the same ZTF and ASAS-SN periodic catalogues that were used in training. Each light curve has a $A/\bar{\sigma} > 2$ (where 'A' is the amplitude calculated as the difference between the 1% and 99% percentile after sigma clipping and '$\bar{\sigma}$' is calculated as the

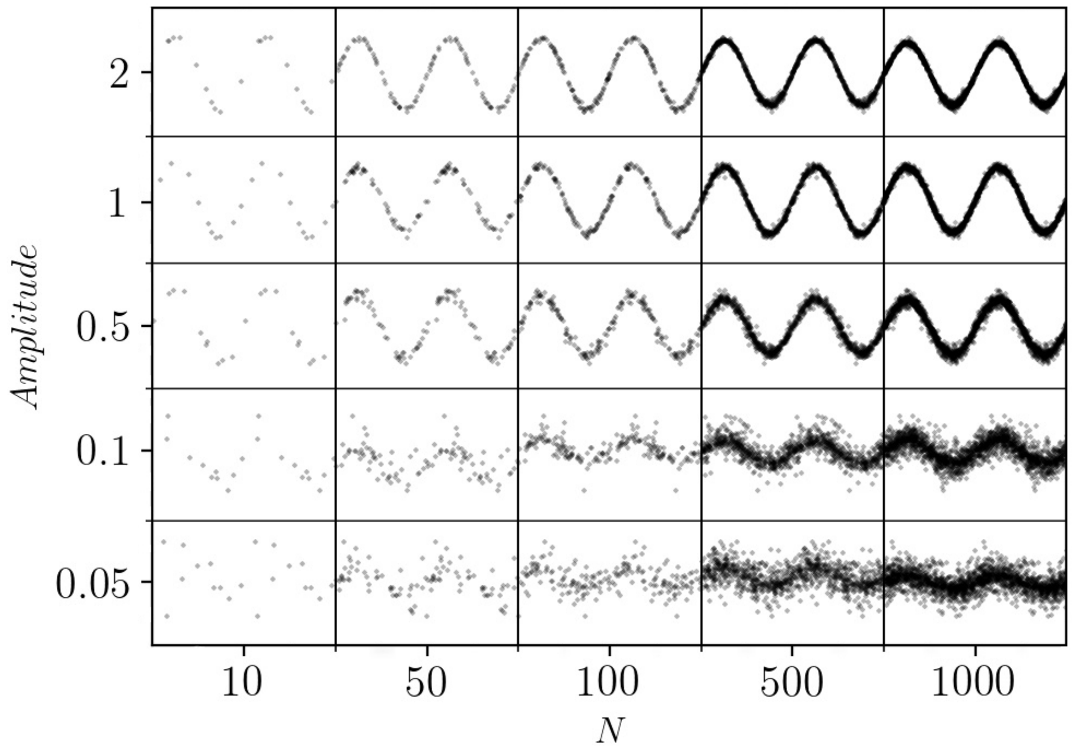


FIGURE 2.4: Examples of a synthetic sinusoidal light curve varying through the number of data points in the light curve on the x-axis and the amplitude of the light curve in the y-axis. The median magnitude error for each point was 0.1.

median value of the magnitude error.) The manual classification of the real light curves involved selecting phase-folded periodic variables by eye. This was independently repeated multiple times by three astronomers to ensure reliability. All of the astronomers agreed on classification. Any ambiguous light curves were removed from the set. Two synthetic data sets were also constructed via the method described in Section 2.3. The data set 'Variable N' was generated as 80 000 identical synthetic light curves with only the number of data points per light curve varied ($10 < \mathrm{N} < 600$). A median SNR ($A/\bar{\sigma}$) of 10 was generated for each of these. The data set 'Variable SNR' was generated as 80 000 identical synthetic light curves with only the signal-to-noise ratio varied. For each light curve in the Variable SNR data set, there were 200 data points used. Figure 2.4 exemplifies both 'Variable N' and 'Variable SNR' on the x and y-axes, respectively. The four types of synthetic variables used were evenly split for both of the synthetic data sets.

Source	NN FAP	Baluev
Real	0.99193	0.95245
Variable SNR	0.99808	0.97843
Variable N	0.99703	0.97393

TABLE 2.1: Showing the AUC for each data set and method.

2.4 Experimental results from RNN

To quantify the performance of the NN FAP we can test its ability as a binary classifier and compare it to the commonly used Baluev method. We use the generalised Lomb-Scargle periodogram along with its associated FAP as described by Zechmeister and Kürster (2009) for our calculations of the Baluev FAP. The Baluev FAP typically lies in a range between unity and 10^{-200} and so the y-axis of the Baluev FAP plots have been shown as both linear and logarithmic scaling.

2.4.1 Performance Measurements

The Receiver Operator Characteristic (ROC) curve is used to measure the capability of a binary classifier as the threshold for classification is varied. An idealised binary classifier will have a threshold at which the *sensitivity* and *specificity* are equal to 1.

Figure 2.5 shows the true positive rate (otherwise known as the sensitivity) versus the false positive rate (otherwise known as 1 - specificity). Equation (2.5) shows more clearly how sensitivity and specificity are defined (where TP and TN are True Positive and Negative respectively. FP and FN are False Positive and Negative respectively.)

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad \text{Specificity} = \frac{TN}{TN+FP} \tag{2.5}$$

The Area Under the ROC Curve (AUC) can be calculated as an evaluation metric for a binary classifier. Table 2.1 shows the AUC for each tested data set. This shows that the NN FAP method has a larger AUC for each data set than the Baluev method. This indicates that the NN FAP method performs better in each test. However, the AUC metric does not tell the whole story, as discussed below.

Figure 2.6 shows the median NN FAP as a function of N and $A/\bar{\sigma}$. This was calculated for 80,000 synthetic sinusoidal periodic light curves (i.e. generated with equation 2.4). The calculations were performed with a range of $3 <$ N < 100 and $0.1 < A/\bar{\sigma} < 2.1$. We can see that this method appears reliable provided $A/\bar{\sigma} >, 1.5$ and N > 40. We note that a median value does not reveal occasional failures and we suggest a limit of N > 50 for greater reliability, based

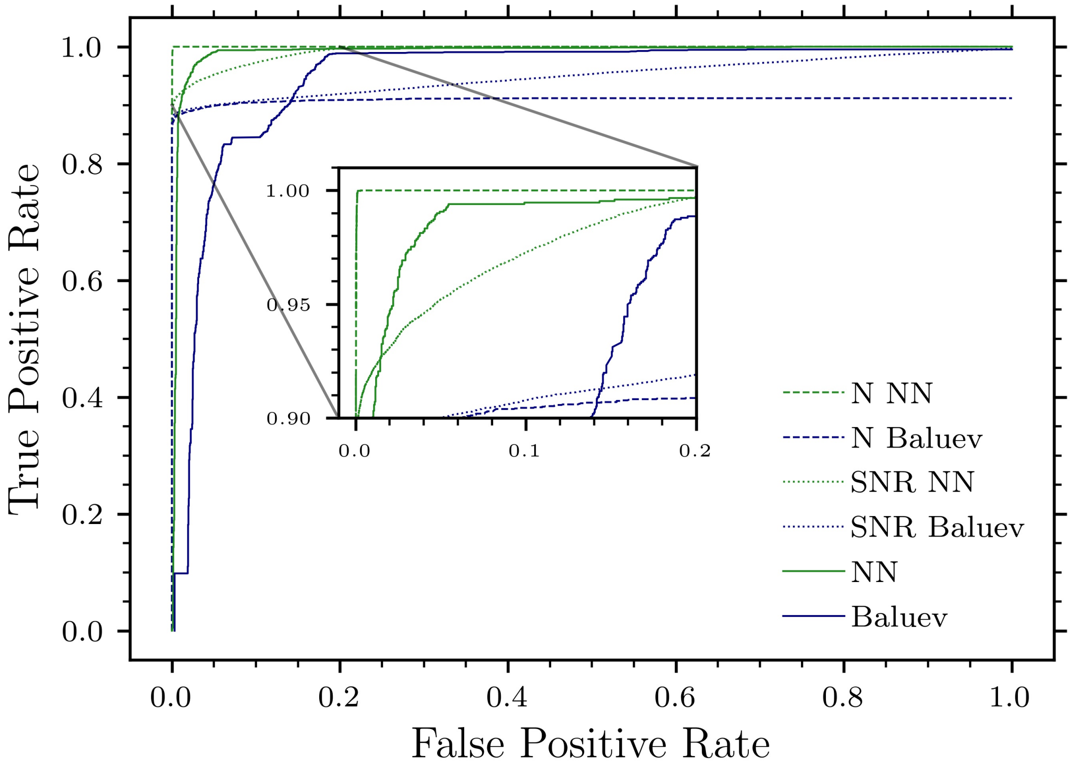


FIGURE 2.5: Showing the ROC Curve for the neural network and Baluev methods as a binary classifier. **Solid line** real data set classified by eye. **Dashed line** synthetic data set where the number of measurements was varied (Figure 2.8). **Dotted line** a synthetic data set where the SNR was varied (Figure 2.9).

on the results in section 2.4.2 and Figure 2.8. A small amplitude with respect to the uncertainty is likely to give false negatives whereas a small number of measurements is likely to give false positives.

We also randomly selected 1000 eclipsing binary stars from the VIVACE catalogue (Molnar et al., 2022). This catalogue was generated from the same VVV data that this model was trained on. All light curves were independently verified as eclipsing binary for this test. We construct a periodogram with both Lomb-Scargle and PDM and choose whichever period produced the lowest NN FAP. We find that 997 of the 1000 were identified as periodic with a FAP < 0.1. The three light curves which failed to be identified each featured a FAP > 0.6. In each of these three light curves there was a significantly shorter transit time paired with N < 60 measurements. The identification of the correct periodicity can be an issue when a light curve can look periodic when phase folded at multiple different periods. If an eclipsing binary features a similar size and shape for each eclipse then the NN FAP can erroneously be assigned half the true period as the two dips in the light curve are likely to be undifferentiable in the phase fold. This can be problematic for equal mass eclipsing binary systems.

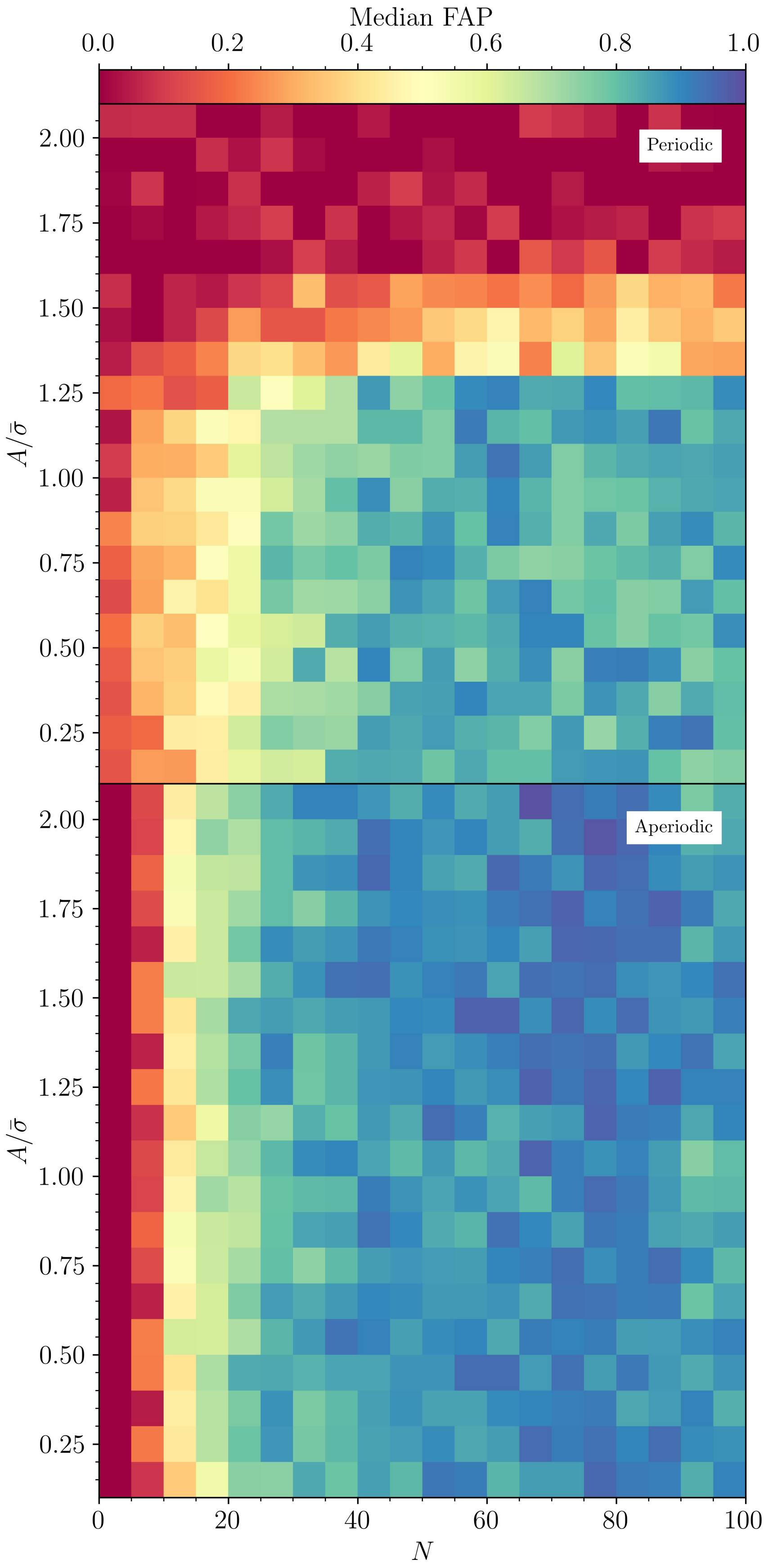


FIGURE 2.6: Showing the FAP calculated for synthetic light curves as a function of the number of data points 'N' in the x-axis and $A/\bar{\sigma}$ in the y-axis

2.4.2 FAP vs N

The number of measurements used to constitute a light curve can vary by orders of magnitude dependent on the survey. Surveys such as Kepler and TESS feature highly sampled light curves which should not pose an issue to any FAP technique. However, this is not always the case and many surveys feature light curves with fewer than 100 measurements. Figure 2.7 shows how the Baluev FAP and the NN FAP vary as a function of the number of measurements 'N' for the synthetic light curve described in section 2.3.4. The NN FAP does not produce any significant number of false negatives as the number of measurements decreases to 10. The Baluev FAP has a clear trend as a function of N and starts to increase to a problematic range of values as N approaches $\sim$50 measurements. It can also be seen that the Baluev FAP has a dependency on the shape of the light curve with more sinusoidal light curves assigned a lower FAP compared to more complex light curves such as eclipsing binary shapes. This is an issue as it can lead to incorrect conclusions on the demographics of variable stars.

Again using the synthetic light curves described in section 2.3.4, in Figure 2.8 (bottom panel) we show that the NN FAP sometimes falls to low values for aperiodic light curves as $N < 50$, potentially leading to false positive classifications. These false positives arise as any variable light curve with a small number of points will more easily represent a periodic light curve at a given phase fold. Caution should be taken with this method when searching for periodic variables with fewer than 50 measurements. By contrast, the Baluev FAP does not suffer from this problem but Figure 2.8 (top panel) shows that it is more likely to assign false negatives to periodic light curves within the same range.

It is not possible to define a threshold for either method which we can use to perfectly separate the periodic and aperiodic light curves. Such a threshold must be set by the user depending on preference regarding completeness and purity. The periodic light curves shown in Figure 2.8 have a maximum NN FAP of 0.791 but the minimum NN FAP for aperiodic light curves is 0.01. The Baluev FAP has a maximum value of 0.015 for periodic light curves but a minimum value of 1.197×10^{-15} for the aperiodic light curves. The Baluev FAP values for periodic and aperiodic light curves overlap despite never approaching 1. The median Baluev FAP for the aperiodic light curves when $N \leq 100$ is 0.0012 and when $N \leq 50$ it is 0.0005. Using the widely adopted criterion for the Baluev FAP of $log_{10}(FAP) < -2$ (Koeltzsch et al., 2009; Herbst et al., 2000; Chen et al., 2020; Botan et al., 2021) yields misidentification of only four of the synthetic aperiodic light curves plotted in Figure 2.8 as periodic, while incorrectly categorising 13,849 (46.4%) aperiodic stars as periodic. This indicates that a lower threshold is more suitable for our synthetic light curves. In Molnar et al. (2022) a Baluev FAP selection of $log_{10}(FAP) < -10$ was used to define a reliable but incomplete set of VVV light curves for training. If we were to use that cut for this data we would misidentify 1152 (3.86%) periodic light curves as aperiodic

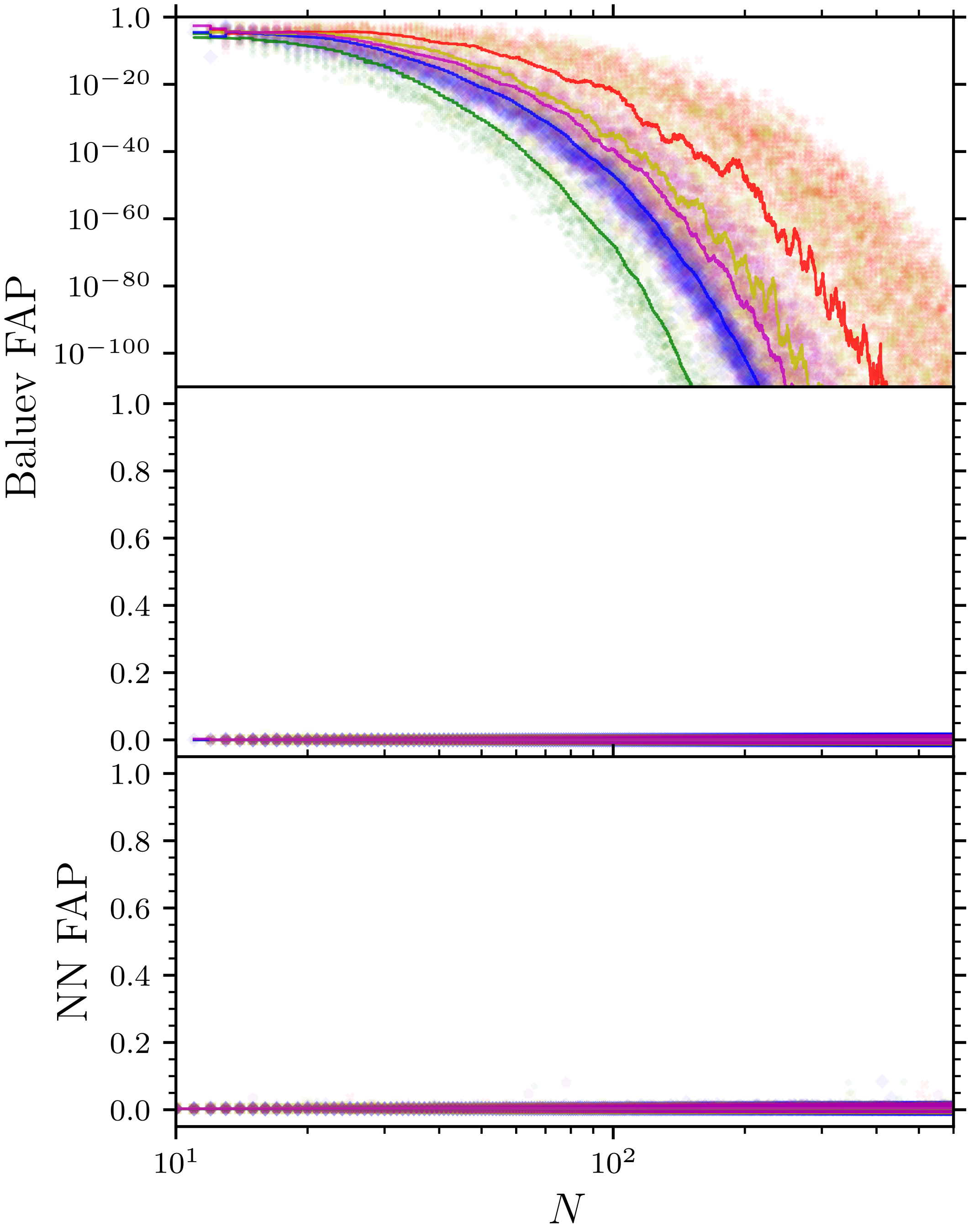


FIGURE 2.7: Both false alarm probabilities versus the number of measurements in the synthetic periodic light curve described in section 2.3.4 The colours and markers correlate to those shown in Figure 2.3. **Top** : Baluev FAP versus N. **Bottom** : NN FAP versus N.

and 502 (1.68%) aperiodic light curves as periodic. If we use a NN FAP of 0.15 we misidentify 1 periodic light curve as aperiodic and 400 (1.34%) aperiodic light curves as periodic.

False positives will arise, or not, depending on the FAP threshold value that is adopted. The NN FAP method performs very well in the AUC test for Variable N (see Table 2.1) because, even where aperiodic light curves have a low FAP, the periodic light curves have even lower FAP values. This allows the binary classifier to be successful, in principle, if the threshold FAP could be ideally selected. However, in practice, this will rarely be possible.

2.4.3 FAP vs Amplitude

The signal-to-noise of a light curve is a common source of erroneous periodicity classification. Periodic variable stars can host a range of amplitudes depending on the source of variability. As such, it is not uncommon to investigate variable stars whose variability is similar to, or below, the photometric uncertainty.

It can be seen in Figure 2.9 that both the NN and Baluev FAP feature a dependency on $A/\bar{\sigma}$. Both the NN FAP and the Baluev FAP suffer from false negative rates as $A/\bar{\sigma} \rightarrow 1.5$. Again, it can be seen that the NN FAP does not suffer from a structure-dependent FAP, unlike the Baluev FAP. This is not surprising as the Lomb-Scargle method, to which the Baluev FAP is applied, is effectively a sinusoidal fitting method and hence will feature such structure-based dependencies.

The median Baluev FAP for the periodic light curves when $A/\bar{\sigma} \leq 1.25$ is 0.011 and 0.020 for aperiodic sources. The NN FAP at the same $A/\bar{\sigma}$ has a median value of 0.959 for periodic sources and 0.998 for aperiodic sources. Both methods feature a significant level of confusion at such a low $A/\bar{\sigma}$ but they do so at different absolute values, the Baluev FAP rarely features values larger than 0.1. If we use the same value of $log_{10}(FAP) < -10$ from Molnar et al. (2022) for the Baluev FAP we misidentify 2028 (3.86%) periodic light curves as aperiodic. If we use a NN FAP of 0.15 we misidentify 1996 (3.79%) periodic light curve as aperiodic. Neither method misidentifies any aperiodic sources as periodic sources with this selection.

2.5 Testing with other surveys

Our proposed method of calculating a FAP is universal and independent of the method of period detection. We can also show that the NN FAP method can be applied to data that is not drawn from the same distribution as the training data. Figure 2.11 shows periodic and aperiodic variable stars in the CRTS (Drake et al., 2009). Figure 2.12 shows them for ZTF. Figure 2.13 shows them

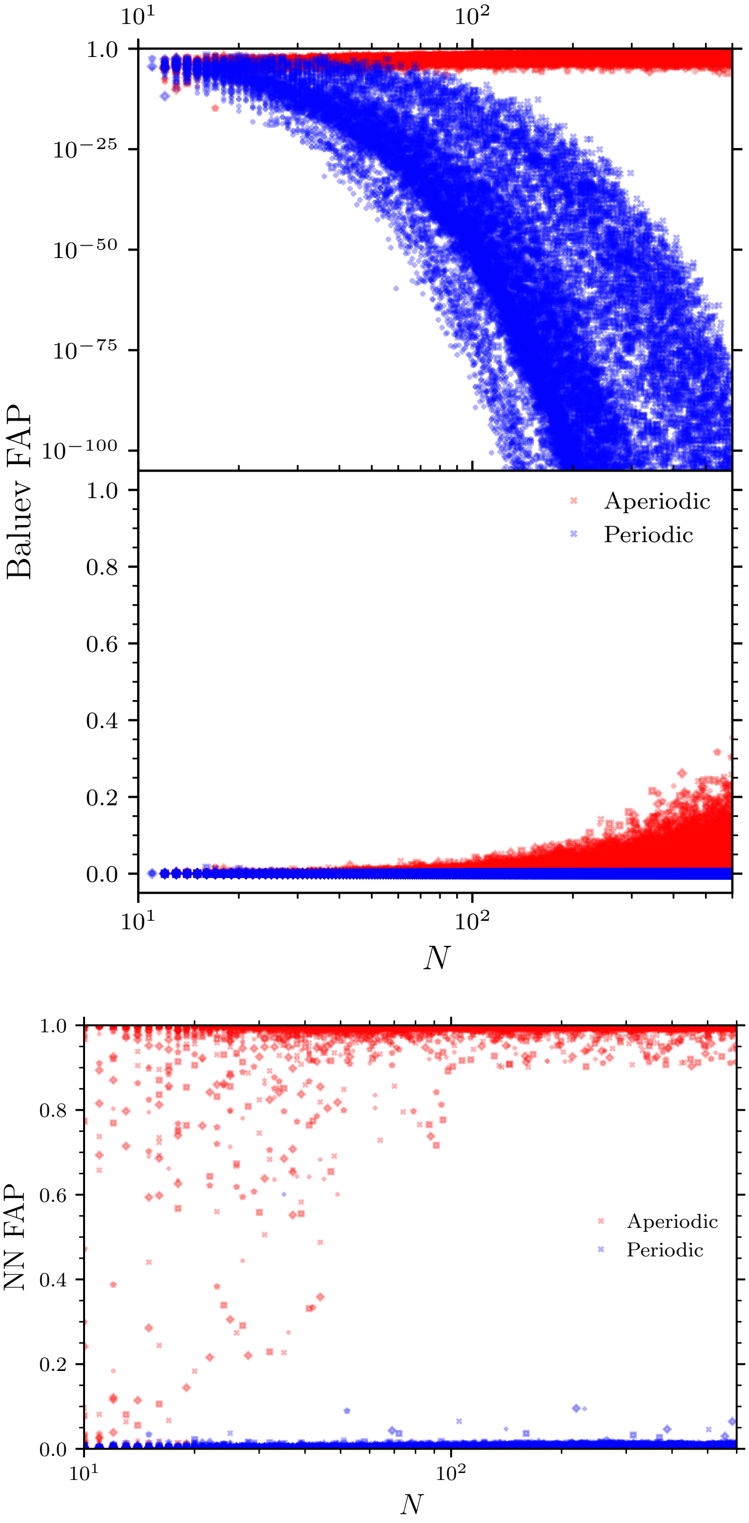


FIGURE 2.8: Both false alarm probabilities versus the number of measurements in the synthetic light curves described in section 2.3.4. The red points show the FAP assigned to aperiodic light curves and the blue shows periodic light curves. It can be seen that the Baluev FAP is more likely to assign false negatives whereas the NN FAP is more likely to assign false positives. The Baluev FAP rarely exceeds 0.1 and never approaches unity. **Top**: Baluev FAP versus N. **Bottom**: NN FAP versus N. Each light curve here featured $A/\bar{\sigma} = 10$. The marker shapes correspond to those shown in Figure 2.3 (i.e. a cross represents 'Type 1' and a plus 'Type 2'...).

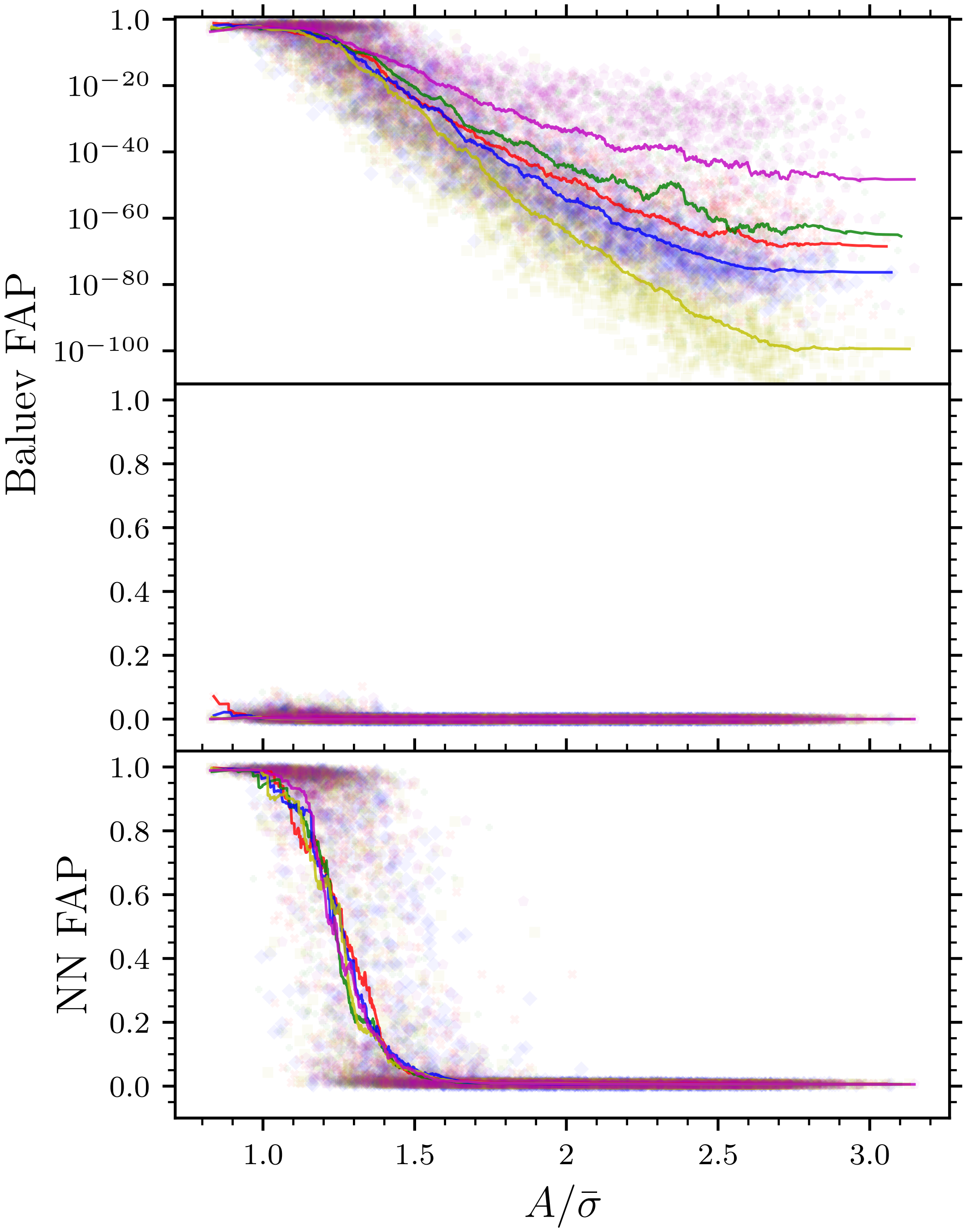


FIGURE 2.9: Both false alarm probabilities versus the amplitude of a synthetic periodic light curve divided by the median average of the photometric uncertainty. The colours and markers are the same as that shown in Figure 2.3. **Top**: Baluev FAP versus $A/\bar{\sigma}$. **Bottom**: NN FAP versus $A/\bar{\sigma}$.

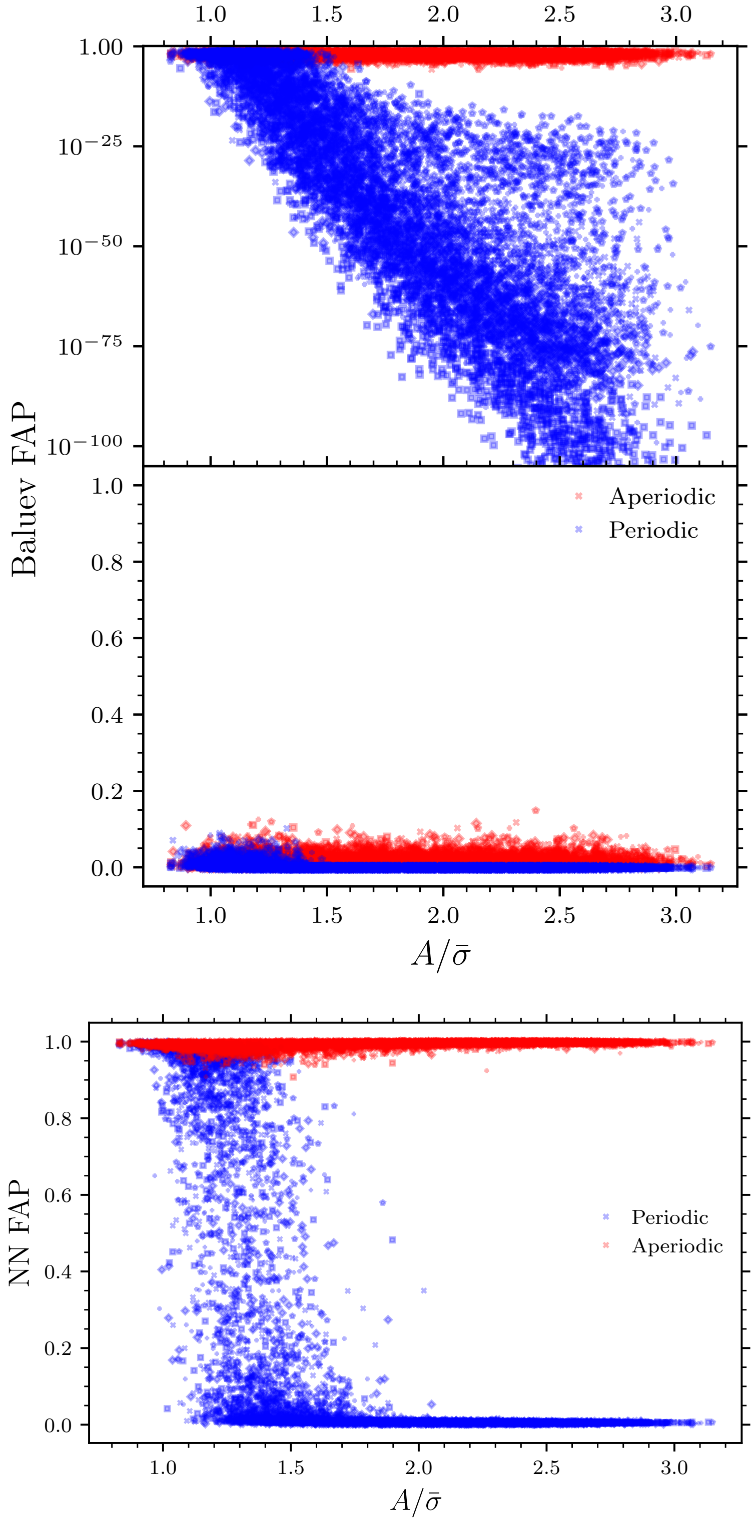


FIGURE 2.10: Both false alarm probabilities versus the amplitude of synthetic periodic and aperiodic light curves divided by their median average of the photometric uncertainty. Both methods show how their reliability begins to fail at $A/\bar{\sigma} \approx 1.5$. **Top** : Baluev FAP versus $A/\bar{\sigma}$. **Bottom** : NN FAP versus $A/\bar{\sigma}$. 200 data points were used for these light curves. The marker shapes correspond to those shown in Figure 2.3.

for kepler. Figure 2.14 shows periodic variables in OGLE (Udalski et al., 2015) data. Each of the sub-plots in these figures show the assigned NN FAP.

The top left panel of figure 2.14 shows the light curves in two filters for the source ”OGLE-BLG-ECL-124368” which appears much more clearly periodic in ‘I’ than ‘V’. The NN FAP reflects this, showing a higher FAP for the ‘V’ band data. The ‘V’ band data does still show a poorly sampled transit at the phase folded period, hence the NN FAP is above 0.5 but below 0.9. The model used to identify these variables was trained as described in section 2.3 with VVV light curves. The periodicity of each of these stars was identified by choosing the period which produced the lowest NN FAP extracted from a PDM periodogram. For both the CRTS and ZTF light curves the Baluev FAP was sufficient for differentiating between aperiodic and periodic variable stars. For three of the Kepler light curves the Lomb-Scargle periodogram incorrectly assigned half of the period with the a low Baluev FAP. One of the Kepler light curves was not identified as periodic by the Baluev FAP (Bottom left panel of periodic variables in figure 2.13). Only one of the OGLE light curves was correctly identified as periodic in both ‘V’ and ‘I’ by the Baluev FAP (Top right panel in figure 2.14) although with a notably different FAP of 9×10^{-60} in ‘V’ and 1×10^{-235} in ‘I’. Each of the other OGLE light curves were either incorrectly given half of the true period or given a Baluev FAP indicative of aperiodic variability. The Lomb-Scargle periodogram also correctly identified the ‘I’ band period of the bottom left panel with a Baluev FAP of 9.51×10^{-141} but failed to extract the correct period for ‘V’ band. The Lomb-Scargle periodogram and Baluev FAP predominantly struggled with more complex eclipsing binary shaped light curves.

2.6 FAP Periodogram

The NN FAP method presented above can be seen as something analogous to a neural network version of the PDM method so we can try to use it as such, i.e. for the construction of a periodogram rather than false alarm probability calculation. We can calculate a FAP for a set of trial periods and the period which returns the lowest FAP should be the correct period. This has the added benefit of generating a periodogram on a universal scale and thus the FAP is given along with the periodogram. Currently, this approach is limited by its computationally demanding nature. Future developments in computing paired with this method being modified for periodogram construction purposes will make this work more practical. Figure 2.15 shows the periodogram constructed for a synthetic light curve (of type 5, Eq. 2.2) with 200 points, a SNR of 2 and a period of 296.4 days. This periodogram took 23 minutes to construct and correctly extracted the correct period (inference was run on 64 CPU cores). This compares to the 0.2 seconds it took for the PDM method to construct the same periodogram and achieve the same results (without a FAP). Both the NN periodogram and the PDM periodogram suffered from

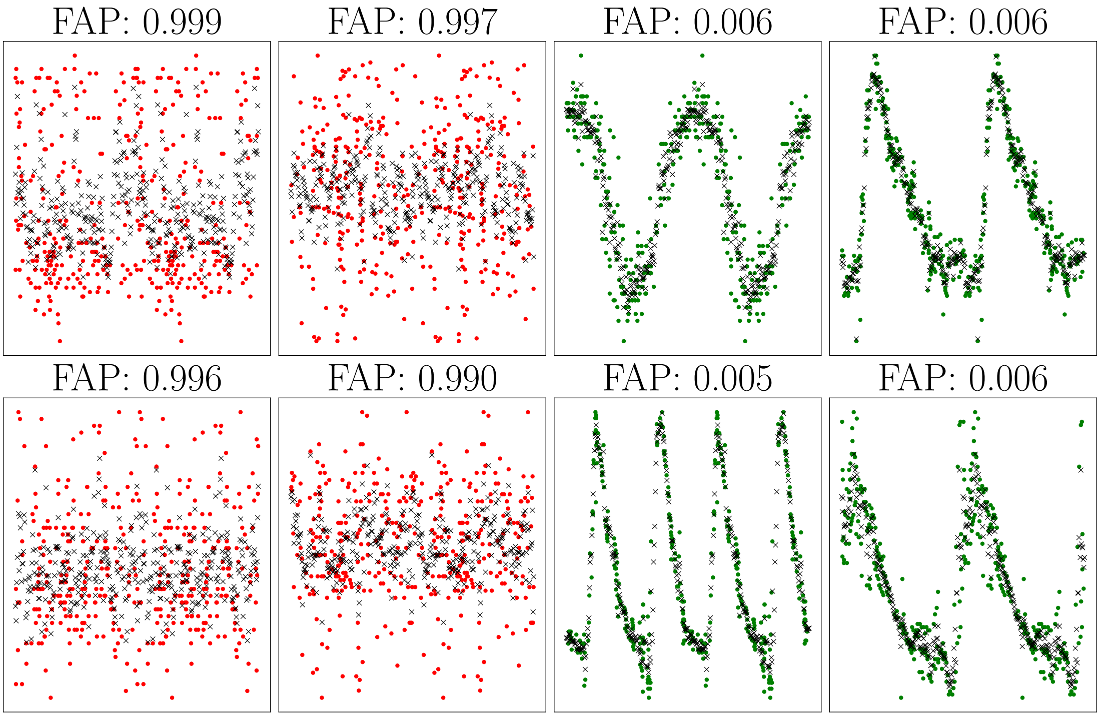


FIGURE 2.11: Randomly selected examples of identified periodic (green, right) and aperiodic (red, left) variable stars found in the CRTS survey. Each subplot displays the assigned NN FAP as its title. The green and red points represent the raw magnitude as a function of phase for the periodic and aperiodic light curves, respectively. The black points represent the KNN interpolated fit to the raw light curve. The Baluev FAP for each of the aperiodic sources was above 2×10^{-5} and the periodic sources were all below 1×10^{-60}.

aliasing at multiples of the true period but both also correctly assigned the true period the largest peak value.

2.7 Conclusions

We have shown that utilising the flexibility afforded by neural networks allows a more robust analysis of light curves. Using synthetic and real data, RNNs can be trained to produce a reliable and universal measure of periodicity.

Our RNN-based method offers an automated, scalable solution that is largely insensitive to the specific shapes of light curves. This allows it to be applied across a diverse array of variable star types, from eclipsing binaries to pulsating stars, without the need for extensive preprocessing or model-specific tuning. It is particularly valuable for surveys like LSST, ZTF, Kepler, and TESS, where the volume of data makes traditional analysis methods impractical. This method can complement existing tools, such as the Baluev FAP, by serving as a secondary and fundamentally different validation mechanism. The flexibility of neural networks also means that our approach

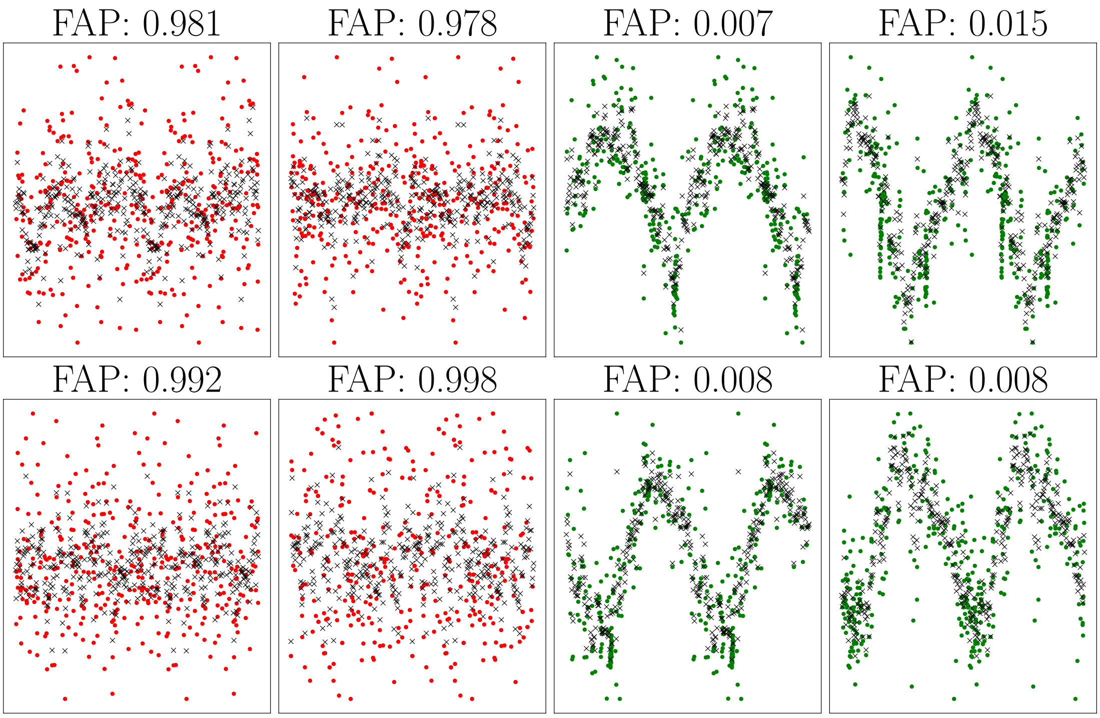


FIGURE 2.12: A random sample of identified periodic (green) and aperiodic (red) variable stars found in the ZTF survey. With each subplot showing the assigned NN FAP as its title. The Baluev FAP for each of the aperiodic sources was above 2×10^{-14} and the periodic sources were all below 1×10^{-51}.

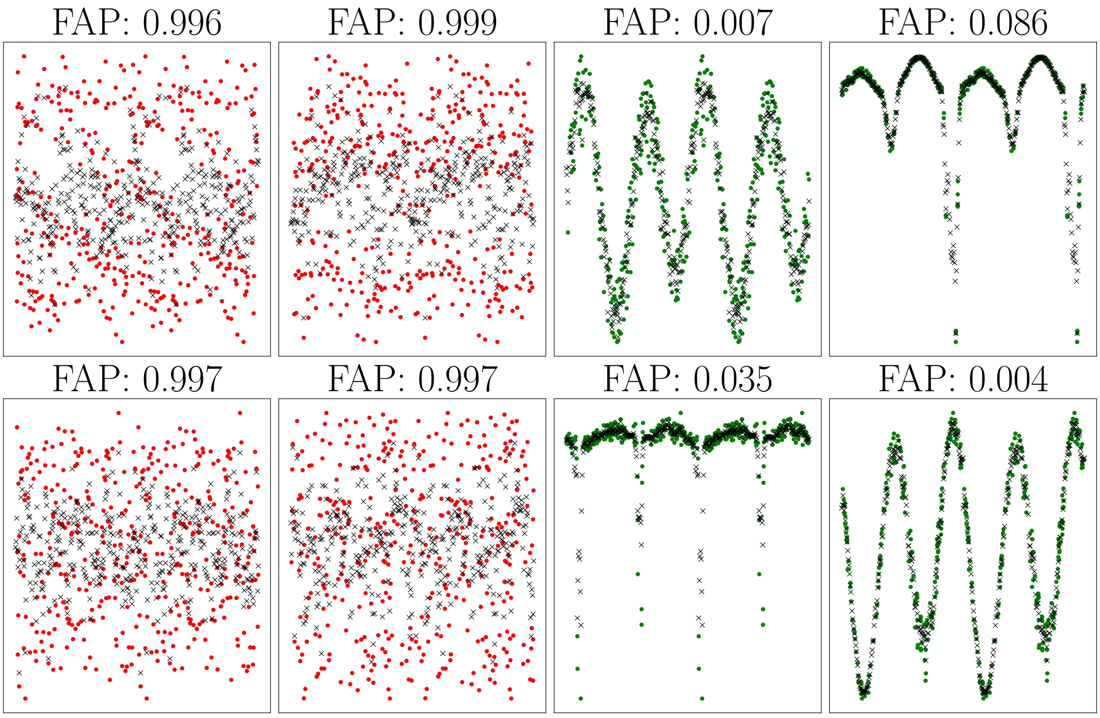


FIGURE 2.13: A random sample of identified periodic (green) and aperiodic (red) variable stars found in the Kepler survey. Each subplot shows the assigned NN FAP as its title. The Baluev FAP for each of the aperiodic sources was above 2×10^{-4}. The bottom left periodic variable has a Baluev FAP of 2.746×10^{-6}. The other periodic sources were all below 1×10^{-43} but each had an incorrect period of half the true period.

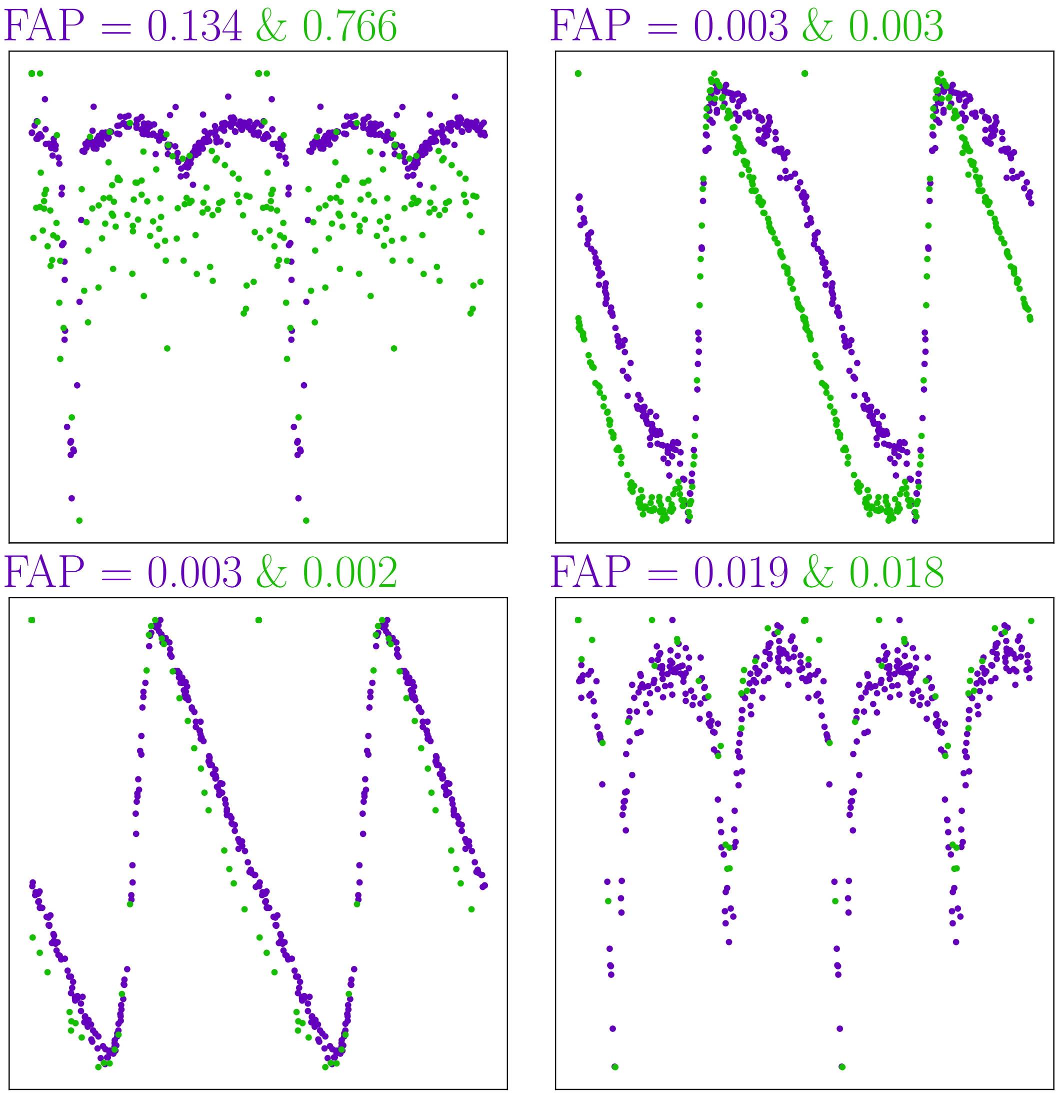


FIGURE 2.14: Periodic variables from the OGLE selection of variable stars. The green points represent the light curve of the star in the 'V' filter and the purple represent the 'I' filter. Three of the stars are identified as periodic in both V and I filters. The top left panel (OGLE-BLG-ECL-124368) was not identified as clearly periodic at any period in 'V' and a higher NN FAP was given (although below that found for the aperiodics in figs. 2.11 to 2.13) From top left to bottom right the Baluev FAPS are 0.967 for 'V' and 0.9 for the incorrect period in 'I', 9×10^{-60} in 'V' and 1×10^{-235} in 'I', 0.016 for the incorrect period in 'V' and 9.51×10^{-141} in 'I', 0.160 in 'V' and 6.9×10^{-65} for the incorrect period in 'I'

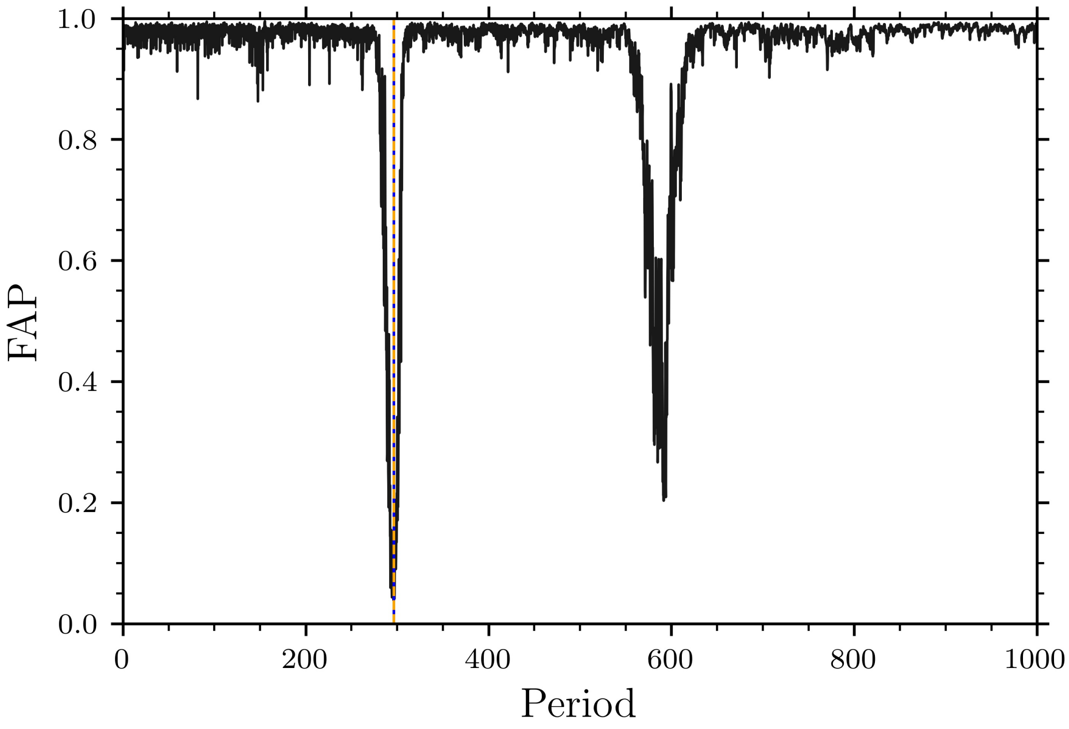


FIGURE 2.15: A periodogram constructed from the NN FAP method. The periodogram took 23 minutes on 64 cores to compute. The correct minima is identified despite the binary-like construction of the periodogram.

can be easily adapted for use in any field where the identification of structure in 2 dimensional data is required.

A study of the parameter space (namely the signal-to-noise ratio and temporal density) demonstrated how and when this method fails in comparison with the commonly used Baluev method. This method remains reliable where $N > 50$ with $A/\bar{\sigma} > 10$ or $A/\bar{\sigma} > 1.5$ with $N \geq 200$. As we analyse the phase-folded light curve and not the periodogram, the NN FAP is independent of the tools used to construct the periodogram. This method is more analogous to a universally scaled PDM and so the network is effectively analysing the structure of the phase-folded light curve. This has further implications for a possible method of period detection that were explored in Section 2.6.

Figure 2.5 and Table 2.1 have shown how this method outperforms the Baluev method for both synthetic and real data. Given a data set for candidate periodic variable stars, this method will provide a more complete search for periodicity, at the expense of occasionally generating more false positives for small N.

We highlight that the most challenging aspect of this method is the data preparation which is outlined in Section 2.3. Care must be given to how the training data is constructed and prepared.

This method is provided both with the ability to retrain on different data sets as well as pretrained with the data described above. We expect the method to be fully functional in its pretrained state within the parameters outlined in this paper. Conversely, this is not the case for the network's architecture which was shown by ablative testing to be relatively inconsequential to the performance.

Chapter 3

PeRiodic Infrared Milky-way VVV Star-catalogue : PRIMVS

3.1 Abstract

We present the PeRiodic Infrared Milky-way VVV Star-catalogue - 'PRIMVS'. We utilise the VVV survey's unique depth and breadth to investigate the variability of astronomical sources within the Galactic bulge and disk. There is a focus on an unbiased and complete identification and classification of periodic variable stars. Employing internal metrics from the VIRAC table for initial selection, we meticulously clean and preprocess light curves to increase reliability and completeness. Care has been taken to address photometric contamination and other sources of uncertainty.

Our approach includes constructing periodograms using Lomb-Scargle, Phase Dispersion Minimisation, Conditional Entropy, and Gaussian Processes to ascertain periodicity. These techniques are used in concert with a novel FAP method (see chapter 2).

This above process allowed us to curate a catalogue of 86,507,172 candidate variable sources.

Machine learning techniques, particularly decision trees and autoencoders, facilitated the initial steps in classification of a significant portion of these sources.

3.2 Introduction

The PeRiodic Infrared Milky-way VVV Star-catalogue - 'PRIMVS' aims to provide a thorough and reliable catalogue of all periodic variable stars present within the VVV survey. The identification of these variables is achieved by the use of parameters present in the VIRAC (Smith et al., 2018) database followed by a variability based selection after light curve cleaning. The compute power of the University of Hertfordshire cluster was heavily utilised for both parallelisation (with a high core count of 128) and the use of GPUS. Care was taken to ensure a minimal amount of quasi-periodic, and otherwise difficult to detect, periodic variables were missed. A Quasi-periodic source is a source whose periodicity is irregular. This irregular behaviour can be caused by an aperiodic change in: period, amplitude, average magnitude or some combination of these. Many statistical measures separate from the identification of a period were made (section 3.6). These statistics serve to provide a full picture of the certainty and reliability of an extracted period and to produce astronomical information, allowing further identification of the source

3.3 Candidate Selection

Due to the uniqueness of the VVV survey, all sources selected for analysis were done so exclusively using internal metrics. An initial selection is made using the variability metrics found in the VIRAC table. These selections are highlighted in table 3.1

Ks detections > 50	Ensure we have at least 50 measurements
Ks detections $> 0.6\,Ks$ observations	Ensure the source is detected at least 60% of the time
$\sigma_{Ks} > 0.01$	Ensure some variability
$\sigma_{Ks}/Ks\_ivw\_err\_mag > 4$	Ensure variability is above some measure of noise

TABLE 3.1: Variability selections performed on VIRAC metadata, prior to light curve cleaning. Where 'σ_{Ks}' is the standard deviation and '$Ks\_ivw\_err\_mag$' is the inverse variance weighted error

These are relatively loose selections aimed at completeness. Due to the high amount of unreliable measurements (as much as 60% in crowded regions) in VIRAC light curves we can't fully rely on variability metrics calculated from the raw light curve. After the initial VIRAC variability selection, the light curve is retrieved and cleaned (see section 3.4). A second check for variability is then made, selections for which are shown in table 3.2.

K_s detections > 50	Reaffirm we have at least 50 measurements after cleaning
Ks error < 0.5	Ensure we have sensible uncertainty
$KsQ_{99} - KsQ_{01} > 0.1$	Ensure there is a minimum of 0.1 mag variability
$KsQ_{75} - KsQ_{25} > 2\text{median}(Ks\text{error})$	Ensure inter-quartile variability is above twice the uncertainty

TABLE 3.2: Variability selections performed after cleaning the light curve

After this selection the light curve is processed as described in section 3.5. Figure 3.1 outlines the process for selecting variables in the VVV data. After selection we are left with 86,507,172 candidate variable sources.

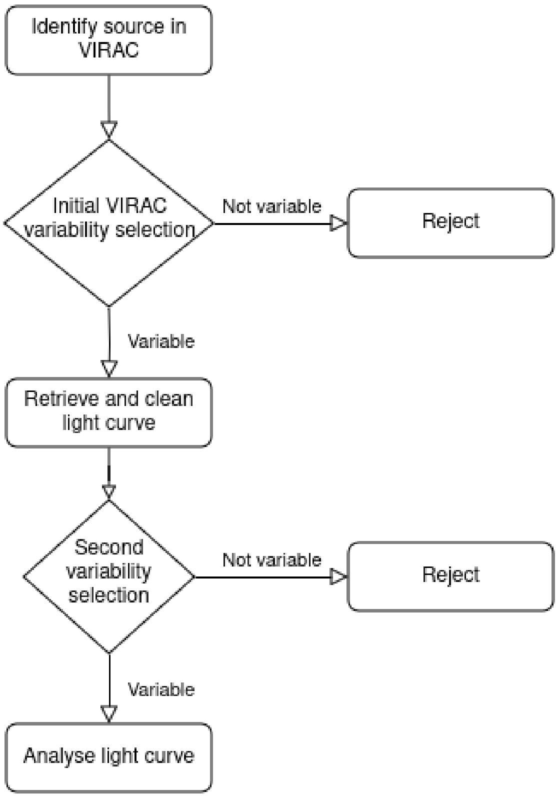


FIGURE 3.1: Flowchart showing the selection process for astronomical sources based on their variability.

3.4 Light Curve Preprocessing

From a practical standpoint, most period-finding methods are relatively simplistic (irrespective of their mathematical complexity). Hence, a large portion of the robustness we achieve in our analysis comes from thoroughly pre-processing a light curve such that a period-finding method will have its effectiveness maximised. This process involves first cleaning the light curve so every measurement is as reliable as possible, and then modifying the light curve, allowing for a more accurate analysis. Due to the depth and survey area of the VVV survey, photometric contamination is a common occurrence.

VIRAC provides multiple metrics of reliability for each measurement in a light curve. Figure 3.2 shows the 'ast_res_chiq' vs 'chi' of a light curve with the dashed line signifying the selection cuts used. Where 'chi' is the DoPhot Chi parameter, representing the quality of the profile fit and 'ast_res_chisq' represents the quality of the 5 parameter astrometric fit to position, proper motion and parallax. It can be seen that the majority of the measurements cluster below the cuts. These cuts were determined with the intention of removing photometry most commonly affected by photometric contamination. This does not serve to remove bad photometry caused by saturation however. It is likely these will also have higher 'chi' values but we do not need to remove them as they will still contribute to any apparent periodicity. If a star is sufficiently

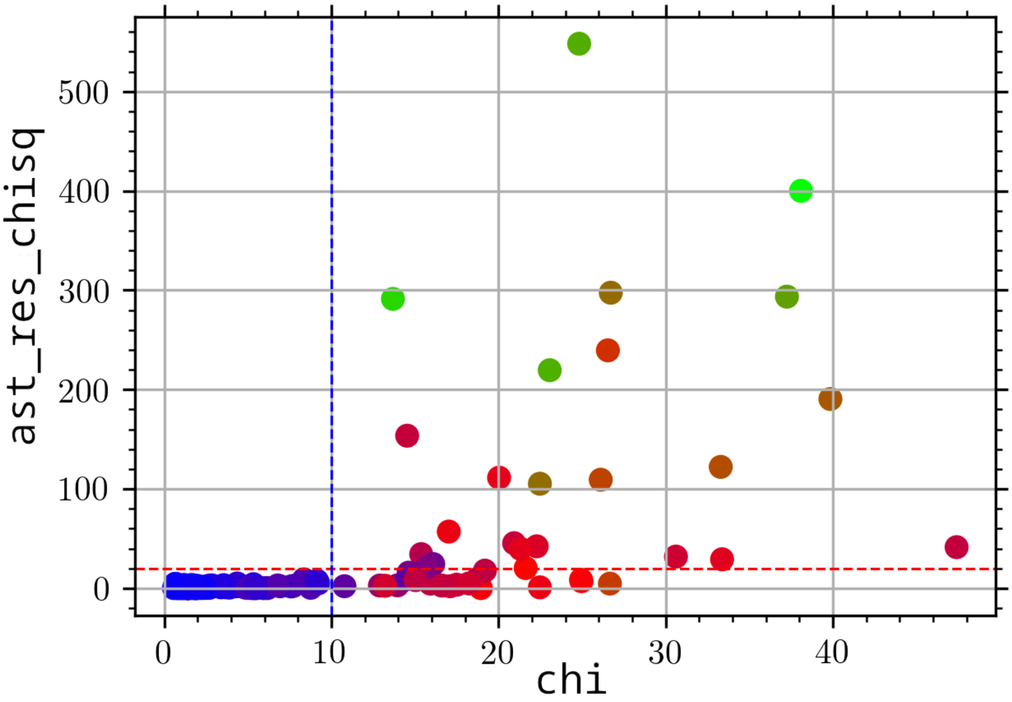


FIGURE 3.2: Showing the range of values taken for a light curve with varying quality of points. Where 'ast_res_chi' and 'chi' are astrometric values taken from the DoPHOT. Colour is proportional to magnitude.

saturated such that the photometric error is problematic, both 'chi' and 'ast_res_chisq' should reflect this and flag the point for removal. There exists a trade off between completeness of the light curves and reliability. Through iterations of the PRIMVS pipeline it has been observed that points with high 'chi' help make the catalogue more complete for bright pulsating stars near the saturation limit. This will enable us to extract a likely period even if amplitudes may be under-estimated. A blanket rejection of points with a magnitude error 0.2 is also applied.

We can also utilise the observing pattern of VVV to both increase the reliability of our data and the photometric certainty. The aforementioned 'paw-print pairs' can be used to check if two points taken close together in time are similar. We do this by ensuring that any pair of data points have similar magnitude errors and are within 2 × $\mathrm{m_{err}}$ of each other. If either are not, both are rejected. After this, any data point within 1 hour of each other with $\mathrm{m_{err}} > 0.1$ are combined by binning 'N' measurements such that $\sigma_{new} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sigma_i$. Any light curve with fewer than 40 measurements after this process is removed from future processing.

After removing erroneous data and combining close points, we move to modifying the remaining data with the goal of making the periodic signal the only photometric variable. A straight line is fitted and subtracted to the light curve, as can be seen in figure 3.4. The linear model is fitted to ensure there are no linear trends throughout the light curve.

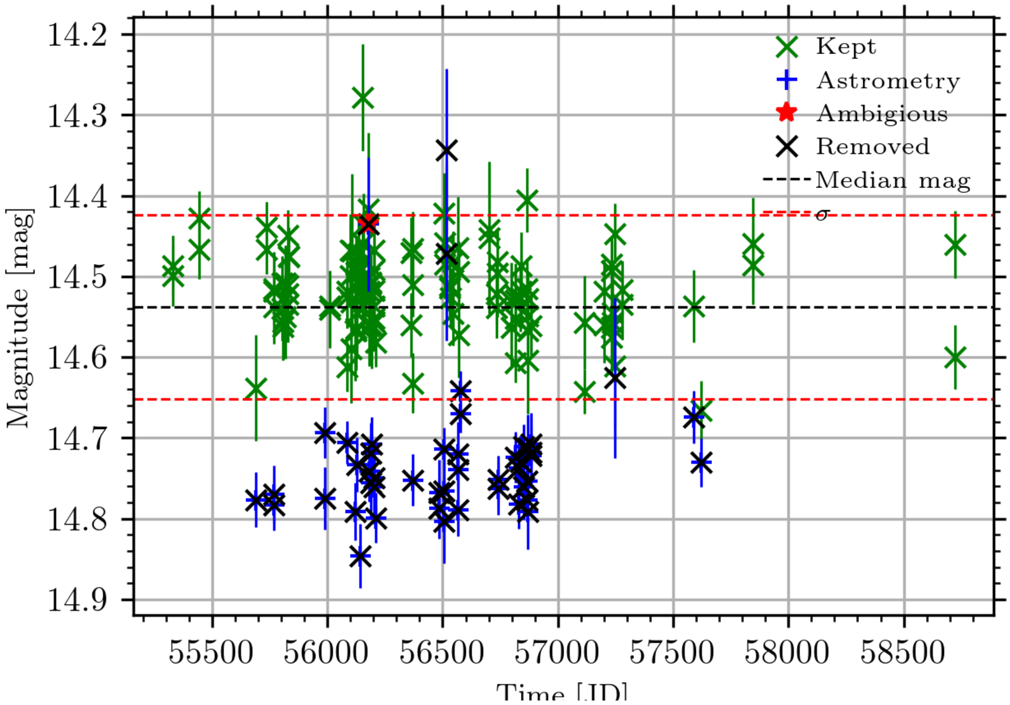


FIGURE 3.3: Showing the points which are removed from a light curve after cleaning. The blue 'Astrometry' points are those that fell outside of the cuts seen in figure 3.2. The red 'Ambiguous' points are determined by a boolean flag which signifies if the source appears blended with a neighbour. The green points are deemed reliable and used for analysis.

AGB stars can feature strong periodic variability on top of a long term variability (Höfner and Olofsson, 2018). While both of these sources of variability provide useful astrophysical information, we are focused on robust period extraction.

Our period-finding methods assume only one source of variability. This can cause an otherwise correctly phase folded light curve to look messy or even incorrect. For most methods, a sufficiently strong linear trend would render all periodic signal extraction virtually impossible.

To subtract the straight-line we first bin the light curve into 10 bins. The weighted median is calculated for each bin and the straight line is fitted to the resulting 10 points. This is done to best capture an overall trend separate to periodic variability whilst also accounting for erroneous outliers. The straight-line is only subtracted if $dm/dt > 2 \times 10^{-4}$ mag/day and $R^2 > 0.2$, where R^2 is the coefficient of determination ($R^2 = 1 - \frac{RSS}{TSS}$, where 'RSS' is the sum of squares of residuals and 'TSS' is the total sum of squares)

The aperiodic form of variability may not be a linear trend and we could fit a higher-order polynomial. However, fitting a higher-order polynomial runs the risk of potentially removing or modifying the periodic source variability, particularly those of longer periods.

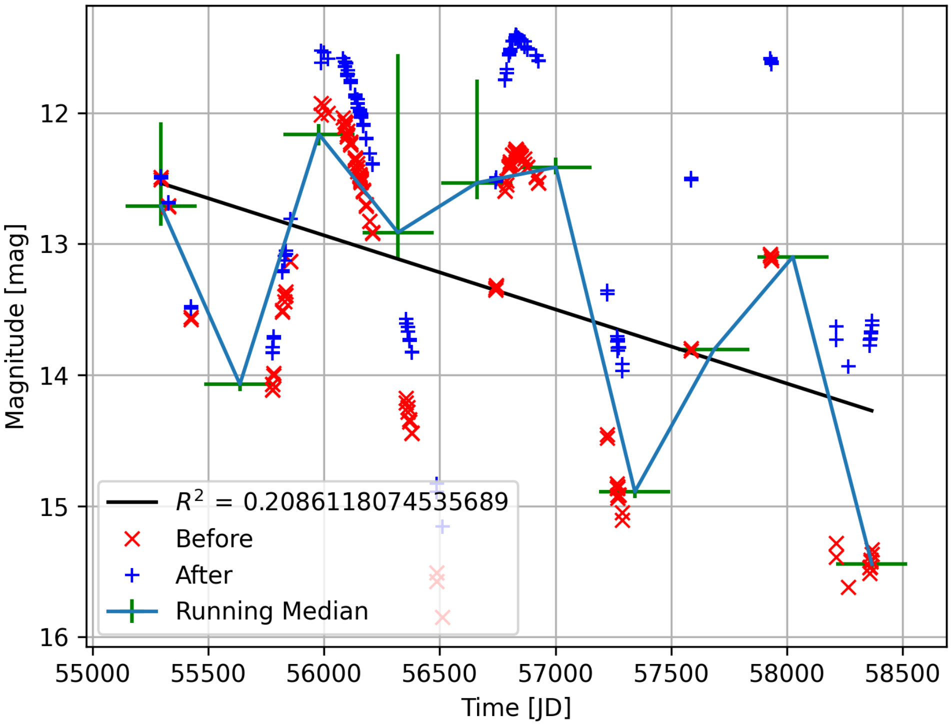


FIGURE 3.4: Showing the measurements before the removal of the linear trend in red crosses along with the fitted black line. The measurements after the removal can be seen as blue plusses '+'.

3.5 Time-series Analysis

After the light curve has been cleaned and prepared for analysis, we can construct a periodogram. We compute periodicity using the Lomb-Scargle (LS), Phase Dispersion Minimisation (PDM), Conditional Entropy (CE) and Gaussian Processes (GP) methods. In an effort to increase the flexibility of light curve analysis, much of the completed testing identifies strengths and weaknesses of each of these methods.

Lomb-Scargle For our implementation of the Lomb-Scargle periodogram, we have used `AstroPy` (Astropy Collaboration et al., 2013). Within the `AstroPy Time Series` package there is a pre-built Lomb-Scargle algorithm (Astropy Collaboration et al., 2013)[1]. We also set '`FIT_MEAN = True`' which enables the Lomb-Scargles generalisation (Zechmeister and Kürster, 2009) for help with smaller data-sets and uneven light curves. It takes an average of ≈ 0.4 seconds to compute a 100000 sample periodogram of a source with 200 data points on 1 core and 1 GB of RAM.

[1] https://docs.astropy.org/en/stable/api/astropy.timeseries.LombScargle.html

Phase Dispersion Minimisation Our version of the PDM periodogram is taken from the original PDM2 source code written in C with edits for efficient interfacing with the python pipeline[2]. We use the updated PDM2 method which has the addition of 'subharmonic sampling'[3]. This uses the PDM window transform to more smoothly bin the phase folded light curve. This allows for clearer differentiation between harmonics of the true period. The python pipeline calls the PDM binary file and passes light curve and periodogram input information via a temporary file. Temporary files are a notable hindrance to the speed of this method as hard disk input/output (IO) operations are orders of magnitude slower than volatile memory operations.

The unreliability of Python's memory management renders temporary files the only appropriate method of sending and receiving data between an external executable. A future version of the PRIMVS pipeline will seek to fix this obvious bottleneck in compute speed. It takes an average of ≈ 0.2 seconds to compute a 100000 sample periodogram of a source with 200 data points on 1 core and 1 GB of RAM.

Conditional Entropy The CE method is a phase folding technique that uses a very similar method to PDM (Graham et al., 2013b). Much like PDM, CE phase folds the light curve for each trial period, then bins the data, and then measures some quantity of the 'scatter' in each of the bins.

This method is fundamentally different from PDM in the way it calculates the 'scatter' of the data points however. Conditional entropy measures the conditional entropy of the phase folded light curve and uses this to quantify the 'scatter' of the data points.

Equation 3.1 describes the conditional entropy $H(m|\phi)$ of a light curve with magnitude 'm' and phase 'ϕ'.

$$H(m|\phi) = \sum_{i,j} p(m_i, \phi_j) ln\left(\frac{p(\phi_j)}{p(m_i, \phi_j)}\right) \tag{3.1}$$

Where $p(m_i, \phi_j)$ is the probability that a data point will occupy the i^{th} magnitude bin of 'm_i' and j^{th} phase bin of 'ϕ_j'. '$p(\phi_j)$' is the probability a data point will occupy the j^{th} bin, which in our case reduces to:

$$p(\phi_j) = \sum_{i} p(m_i, \phi_j) \tag{3.2}$$

Configuring the amount of bins used for the phase axis is crucial in the Conditional Entropy (CE) method, as it directly affects the sensitivity and accuracy of detecting periodic signals. The

[2] `https://www.stellingwerf.com/rfs-bin/index.cgi?action=PageView&id=34`

[3] `https://www.stellingwerf.com/rfs-bin/index.cgi?action=GetDoc&id=21&filenum=1`

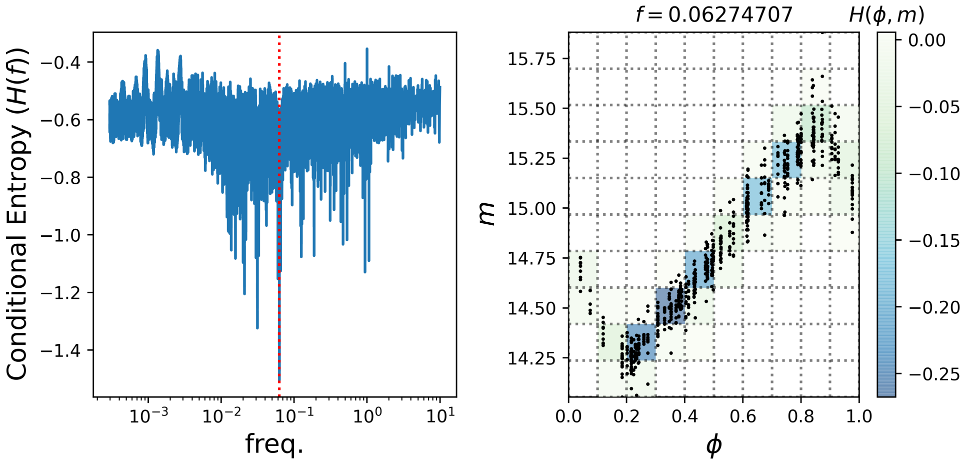


FIGURE 3.5: **Left**: A plot showing conditional entropy as a function of frequency. **Right**: Showing the phase folded light curve that produces the lease total conditional entropy. In this plot we can see each of the bins used as well as the conditional entropy each of them hold.

jackknifing method from Hogg (2008) is used to determine the number of magnitude bins to use for each light curve. However, it would be computationally expensive to calculate the optimal bins for the phase axis in each trial period. We opted for 10 phase bins based on recommendations from Graham et al. (2013b), striking a balance between resolution and noise. Fewer bins may result in a loss of resolution, making it challenging to detect subtle variations in light curves. This will be particularly problematic for sources with complex variability patterns like pulsating stars or eclipsing binaries. Conversely, using too many bins can lead to overfitting, where the conditional entropy becomes dominated by noise rather than genuine signal features. This will result in a noisy representation of the phase distribution. Future work may explore adaptive binning techniques that adjust the number of bins based on the light curve's characteristics, potentially improving sensitivity across different variable star types and increasing the robustness of period detection.

Figure 3.5 shows the periodogram (left) and the phase folded light curve when plotted at the optimal period (right). This figure highlights how this method works as well as the importance of the bins.

We can see from figure 3.5 that if we were to reduce the number of bins, the resolution of the process would effectively decrease as it would be harder to differentiate between small changes in the shape of the light curve. However, if we increase the number of bins too much, the conditional entropy would be dominated by small changes. This results in a noisy representation of the phase distribution.

For our implementation of CE, we have used the python package '`cuvarbase`'. The CE periodograms were not constructed as part of the main three tests. The computational intensity of

CE renders it a GPU bound operation. The University of Hertfordshire High Performance Cluster features 6 GPU nodes, each with at least the computational equivalent of 3 Tesla A100 16GB GPUs. It typically takes ≈ 1 second to recover a period, as opposed to the ≈ 20 minutes for the same process when CPU bound. However, there are only 32 normal cores with 32 GB of RAM for each of these GPU nodes. This is significantly less than the 256 cores and 128 GB of RAM that is used for the combined LS and PDM test. At the time of writing, the CE periodogram is only computed for sources with multiple distinct periods with a low FAP in an attempt to clarify ambiguities. A future version of PRIMVS will be made where significantly more/all of the sources are analysed with CE.

Gaussian Processes A periodic signal observed in astrophysics is rarely a perfect sinusoid. Often periodic signals in this field vary in non-sinusoidal and Quasi-Periodic (QP) ways. To effectively model this behaviour we would ideally have a small number of parameters that are flexible enough to properly describe real astrophysical signals. In Rasmussen and Williams (2006) Gaussian Processes are described as providing a "...principled, practical, probabilistic approach to learning in kernel machines." Gaussian Processes are unique in our comparison of period finding techniques, as their ability to identify a periodic signal is only a product of the kernel used. Through different kernels and different combinations of kernels, Gaussian Processes can model many patterns within data. The flexibility present in Gaussian Processes is from their modelling of the covariant structure of the data, rather than the absolute values of data. This means that a relatively simple kernel is likely to be able to describe the structure of many light curves.

There are many kernels, and combinations thereof, available to use for Gaussian Processes. In Rasmussen and Williams (2006) the Quasi-Periodic kernel is used to measure the concentration of CO_2 on the summit of the Mauna Loa volcano in Hawaii. To achieve this, a product of two basic kernels are used; the squared exponential kernel and the periodic kernel. In Angus et al. (2018) GPs are used to identify the often quasi-periodic nature of stellar rotation periods, the QP kernel is used.

$$k_{i,j} = \mathrm{A}\exp\left[-\frac{(x_i - x_j)^2}{2l^2} - \Gamma^2 \sin^2\left(\frac{\pi(x_i - x_j)}{P}\right)\right] + \sigma^2\delta_{i,j} \tag{3.3}$$

Where '$k_{i,j}$' is the covariance between points 'x_i' and 'x_j'. 'A' is the amplitude factor, scaling the overall covariance. '$\exp\left[-\frac{(x_i-x_j)^2}{2l^2}\right]$' is the radial basis function (RBF) which models the smooth variation in the data. 'l' is the length scale of the RBF kernel, controlling how rapidly the similarity between two points decreases as their distance increases.

'$\exp\left[-\Gamma^2\sin^2\left(\frac{\pi(x_i-x_j)}{P}\right)\right]$' is the periodic component of the kernel, where 'P' is the period, and 'Γ' adjusts the relative importance of the periodic versus RBF component.

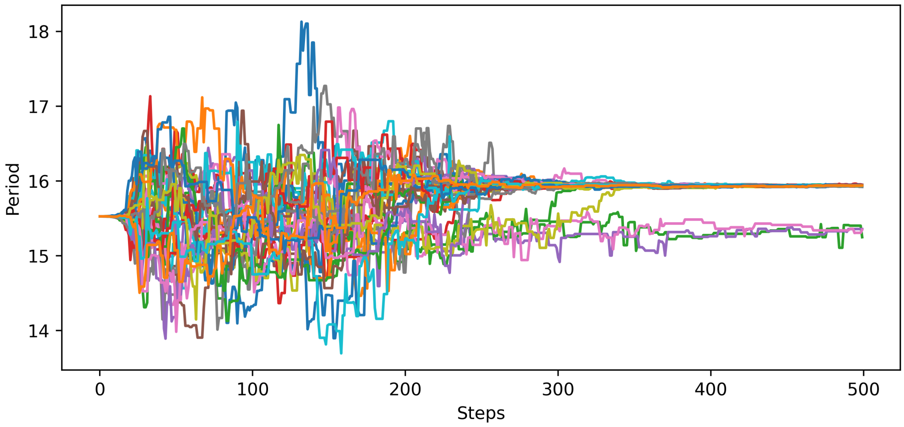


FIGURE 3.6: Showing the path the 32 walkers took in their 500 steps they made within the MCMC process. It can be seen that after ≈ 350 steps the majority of the walkers fall into what is approximately the same value for the period with a few which do not.

'$\sigma^2\delta_{i,j}$' represents the noise term, where 'σ^2' is the variance of the noise, and '$\delta_{i,j}$' is the Kronecker delta function, equal to 1 if $i = j$ (i.e., for the diagonal elements representing the variance at each point) and 0 otherwise. This term tries to account for uncorrelated noise measurements, the presence of a Kronecker delta asserts that each measurements noise is independent.

The GP kernel is minimised with both '`scipy minimise`' and '`emcee`' python packages. The '`scipy minimise`' package uses least-squared regression which can struggle with the number of free parameters in the dataset. This is mostly used to provide an initial position for each of the walkers in the Monte Carlo Markov Chain (MCMC) process whic follows. The MCMC minimisation utilised 32 walkers and 500 steps. Statistical analysis is performed on the last 50 steps of the MCMC minimisation process. Figure 3.6 shows the path the walkers took under MCMC minimisation. It can be seen that a small number of walkers deviate from the majority. Without removing the values these walkers represent, any averages drawn from the total output are at risk of being erroneously shifted by these walkers. In the future, it might be possible to model systems with multiple periods by identifying separate clumps of walkers. For this iteration of the PRIMVS pipeline, the GP is run if the straight line fit has $dm/dt > 2 \times 10^{-4}$ mag/day(from the end of section 3.4), FAP > 0.2, Amplitude > 0.5 and there are more than 100 measurements. This is done to save compute and utilise the GPs ability to identify periodicity with additional trends present. Much like CE, the intention is to analyse most/all of the PRIMVS catalogue with this method.

3.5.1 Period Searches

A periodogram will be constructed from a list of test periods. Period finding methods will take a set of test periods, apply that period to the light curve (either via phase folding or fitting) and

Test #	Period range
Test 1	1 d <P <500 d
Test 2	0.01 d <P <1 d
Test 3	500 d <P <$T_{lc}/2$

TABLE 3.3: Table showing each of the three successive period scans used

then measure some feature of the light curve (goodness of sine fit, binned scatter...). In order to maximise completeness in a blind search for periodicity we must ensure our set of trial periods does not impart a selection bias. It is common practice to search for periodicity linearly in frequency space (Chen et al., 2020) as doing so in period space will disproportionately compute for longer period variable stars (e.g. a periodic variable star will look fine when phase folded at 50.1 days given a true period of 50 days. This is not the case for a star with a period of 5 days which has been phase folded at 5.1 days).

For most completeness we split our period search into three successive searches of 100,000 trial periods, the ranges of which are seen in table 3.3. For each of the three tests we remove any sources with a FAP < 0.1 from future tests to save on compute.

Where '$\frac{T_{lc}}{2}$' is the length of the light curve divided by two, ensuring a minimum of 2 cycles are seen. Figure 3.7 shows the pipeline for the period analysis.

3.5.2 Periodogram

Another important area for ensuring the thorough analysis of a time-series data set is the treatment of the constructed periodogram. For each periodogram, we first exclude known problematic areas, such as those from the lunar, diurnal and yearly cycles. Then the three most significant peaks are identified. This is not always the 3 highest distinct points on the periodogram. To identify a peak we must also find a trough on either side. This stops us from extracting a particularly wide peak twice or extracting a peak which is instead just the edge of the periodogram. It also allows us to characterise the peak width and height.

After each periodogram has its 3 most prominent peaks extracted, their corresponding periods are measured and compared to each other. The FAP which we used allows for the universal comparison of periodicity regardless of the method used to identify the periodicity. This allows us to mitigate a lot of the biases that are exclusive to either method. For example; LS struggles more than PDM with the identification of periodicity for eclipsing binaries (VanderPlas, 2018) and so in cases where LS will fail, PDM should succeed and be recognised by the lower FAP. This method also allows us to account for some of the effects that sampling, aliasing and other perturbers may have in hiding the true period amongst other high peaks. If the true period is

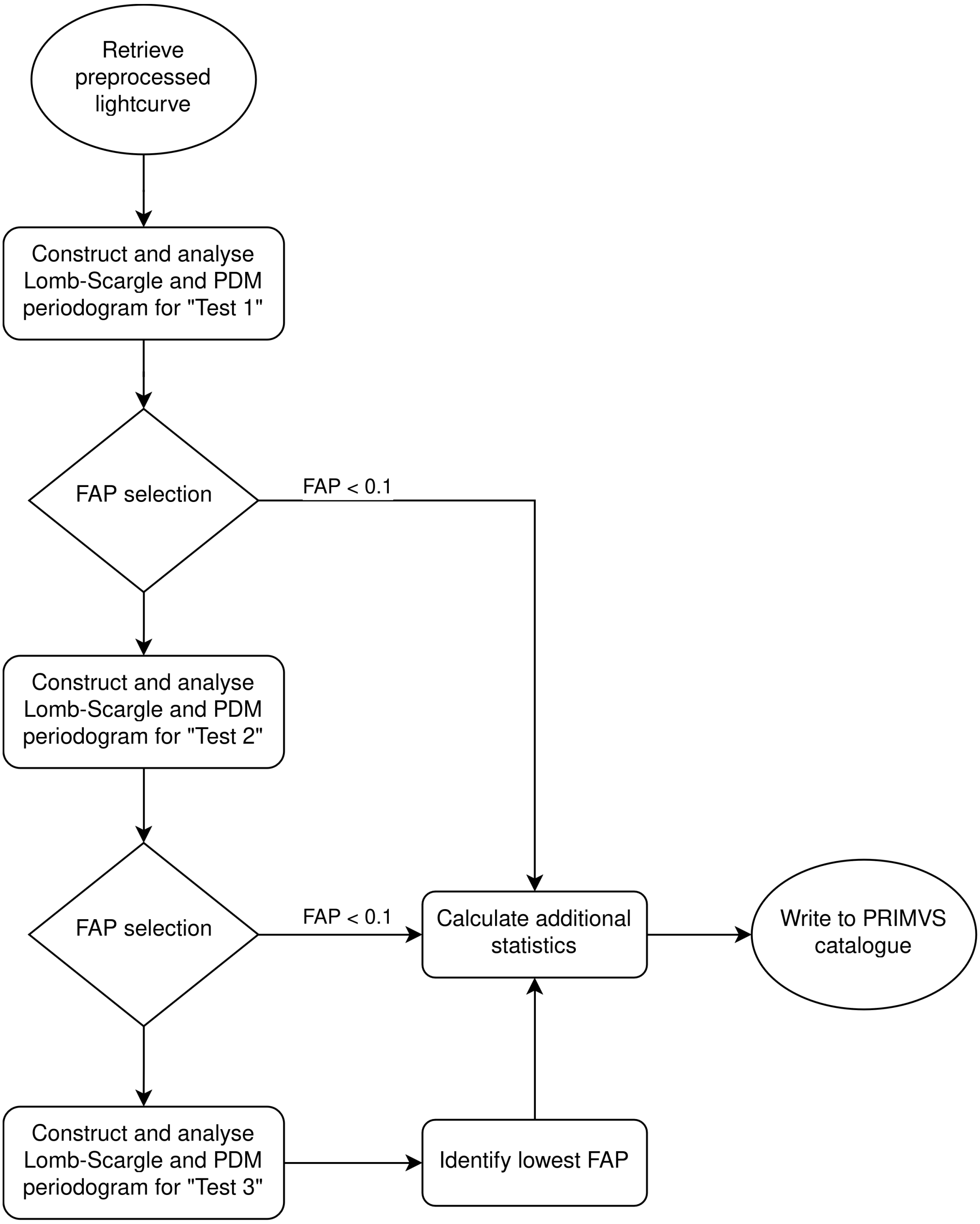

Retrieve preprocessed lightcurve
Construct and analyse Lomb-Scargle and PDM periodogram for "Test 1"
FAP selection
FAP < 0.1
Construct and analyse Lomb-Scargle and PDM periodogram for "Test 2"
FAP selection
FAP < 0.1
Calculate additional statistics
Write to PRIMVS catalogue
Construct and analyse Lomb-Scargle and PDM periodogram for "Test 3"
Identify lowest FAP

FIGURE 3.7: Flowchart showing the pipeline for processing the light curves through each of the three tests seen in table 3.3. Sources identified with a FAP<0.1 are removed from future tests. Each test consists of 100,000 trial periods.

within the top 3 most significant peaks then it should be identified by the neural network FAP (see chapter 2), which operates independently of the periodogram.

3.5.3 Period Alias Check

A common issue with the analysis of periodicity is knowing which alias of the period is correct; that is, if we extract two similarly likely periods at P and $0.5 \times P$ or $2 \times P$, which one is correct?

It is typical to construct a periodogram with a resolution either linear or logarithmic in frequency space (i.e not linear in period space). This is done to sample the frequency space as completely as possible while still being computationally feasible. If we have a source with a period towards the end of our linearly sampled periodogram, it could be possible that we fail to sample at the specific period of the source. However, it could be more likely we sample at half of that period (as the sampling density in period space will increase towards shorter periods). This could erroneously heighten the peak at this period above that of the true period.

While this problem typically hinges on the ability to confidently determine which phase folded period is actually correct (which is often impossible), we can at least ensure a fair test is performed for each trial period. This is achieved by recomputing the periodogram with an increased density of trial periods at previously identified significant peaks. This ensures that we do not miss the true period due to insufficient sampling density. After we have obtained the aforementioned list of unique candidate periods (described in chapter 2) we can recompute the periodogram before calculating a FAP.

At the time of writing, this method computed for all sources with FAP < 0.1 and P > 1000 days (1,028,397 sources). This method is computationally expensive and so was only used on what is expected to be the most secondarily affected sources. The intent is to use this approach of recomputing periodograms for all objects. For the first iteration of PRIMVS, we exclude period ranges that are close to known diurnal and lunar aliases, such as 1 day, 2 days, 29 days, and 60 days. This exclusion applies to these specific period ranges, not to the variables themselves, which might otherwise exhibit periodic behaviour. By avoiding these alias-prone ranges, we aim to reduce the likelihood of misidentifying false periods as genuine. These cuts were selected based on the high susceptibility to observational artefacts within these ranges, which can produce misleading peaks in periodograms. While this approach helps minimise erroneous detections, it also potentially overlooks true periodicities that coincide with these alias ranges. To address this limitation, future iterations will explore the integration of machine learning techniques, similar to those employed by Christy et al. (2023), to better distinguish between genuine and false periods.

The typical readout time for the VIRCAM detectors is around 1 second, with additional overhead for data processing and telescope jitter movements. This results in a timing precision for individual measurements of a few seconds. Given this timing precision, the theoretical minimum period that can be resolved is a few seconds. However, in practice, periods shorter than a few minutes may be challenging to detect reliably due to noise, stochastic sampling, and additional overheads. No uncertainty for time is given in the VIRAC data.

3.6 Further statistics and False Alarm Probability

After the periodograms have been constructed and analysed the FAP is calculated (via the method described in chapter 2). Other statistics, particularly relating to the magnitude distribution, are also calculated.

The following is a description of each of the (groups of) features in the PRIMVS catalogue. The PRIMVS catalogue has two distinct sets of features: light curve features and periodogram features.

mag_n, time_range, mag_avg and magerr_avg Number of points, time range, median magnitude, and median magnitude error in the cleaned light curve (i.e. the light curve that was used for analysis which is different from the raw light curve)

The mean, standard deviation, skew and kurtosis of the light curve is also calculated. The error weighted counterpart for each of those values is also used.

The following representations of magnitude and error will be used for all further feature definitions; m, $\bar{m}$, σ_m as magnitude, mean magnitude and magnitude standard deviation respectively. Magnitude error is denoted as m_{err} with the same variations used for magnitude ($\bar{m}_{err}$, σ_{merr}).

true_period The most likely period '*true_period*' is the potential period which had the lowest FAP (see chapter 2)

best_fap The lowest FAP from each of the extracted potential periods. i.e. the FAP which is obtained from the '*true_period*'

Cody_M Measure of asymmetry '*M*' from Cody et al. (2014).

$$M = \frac{\bar{m}_p - \text{median}(m)}{\sigma} \tag{3.4}$$

where

$$\bar{m}_p = \frac{1}{N_p} \sum_{m_i \in P} m_i \tag{3.5}$$

and

$$P = \{m_i \,|\, m_i > Q_{90}(m) \text{ or } m_i < Q_{10}(m)\} \tag{3.6}$$

where '$Q_{90}(m)$' and '$Q_{10}(m)$' are the 90^{th} and 10^{th} percentiles of the magnitude distribution. 'N_p' is the number of points in the set 'P', and 'σ' is the overall root mean square of the magnitude distribution.

Stetson_K Robust measure of kurtosis '$Stetson\_K$' (Stetson, 1996)

$$Stetson\_K = \frac{1/N \sum_{N}^{i=1} |\delta_i|}{\sqrt{1/N \sum_{N}^{i=1} \delta_i^2}} \tag{3.7}$$

where the relative error 'δ' is defined as

$$\delta = \sqrt{\frac{N}{N-1}} \frac{m - \bar{m}}{m_{err}} \tag{3.8}$$

Where the number of points in the light curve is N.

von Neumann η and η_e The von Neumann variability indices 'η' (Neumann, 1941) and 'η_e' (Kim et al., 2014) was developed as a check for whether successive data points are independent.

$$\eta = \frac{\sum_{i=1}^{N-1} (x_{n+1} - x_n)^2 / (N-1)}{\sigma_m^2} \tag{3.9}$$

However, this assumes we have evenly spaced samples and so we also have

$$\eta_e = \bar{w}(t_{N-1} - t_1)^2 \frac{\sum_{i=1}^{N-1} w_i (m_{i+1} - m_i)^2}{\sigma_m^2 \sum_{i=1}^{N-1} w_i} \tag{3.10}$$

where

$$w_i = \frac{1}{(t_{i+1} - t_i)^2} \tag{3.11}$$

which takes into account the stochastic sampling of our data.

medianBRP The 'median buffer range percentage' (Richards et al., 2011) is the percentage of points within the one tenth of the maximum amplitude.

$$\text{medianBRP} = \frac{|S|}{N} \tag{3.12}$$

where '$|S|$' is the number of points within '$A/10$' of the median magnitude '$\bar{m}$'. '$A/10$' is the amplitude divided by 10.

$$S = \{x \in m | \bar{m} - A/10 < x < \bar{m} + A/10\} \tag{3.13}$$

range_cum_sum The range of a cumulative sum (Ellaway, 1978). The R_{cs} should tend to 0 for symmetric distributions.

$$R_{cs} = \max S - \min S \tag{3.14}$$

where

$$s = \frac{1}{N\sigma_m} \sum_{i=1}^{N} (m_i - \bar{m}) \tag{3.15}$$

max_slope The maximum gradient between two points in the cleaned light curve.

$$\text{Max slope} = \max_{1 \leq i < N} \left|\frac{m_{i+1} - m_i}{t_{i+1} - t_i}\right| \tag{3.16}$$

MAD The median absolute deviation 'MAD' of the magnitude distribution.

$$\text{MAD} = \text{median}(|m - \text{median}(m)|) \tag{3.17}$$

mean_var The mean variance '*mean_var*' can be used as a simple indication of variability.

$$mean\_var = \frac{\sigma}{\bar{m}} \tag{3.18}$$

percent_amp The percentage amplitude '*percent_amp*' is the largest percentage difference from the median value.

$$percent\_amp = \frac{\max(m_i - median(mag))}{median(mag)} \tag{3.19}$$

roms The Robust Median Statistic '*roms*' is a metric of variability.

$$roms = \frac{\sum_{i=1}^{N} |m_i - median(m)|/m_{err,i}}{N - 1} \tag{3.20}$$

ptop_var The peak-to-peak variability '*ptop_var*' is effectively the weighted percentage amplitude.

$$ptop\_var = \frac{\max(m - m_{err}) - \min(m - m_{err})}{\max(m - m_{err}) + \min(m - m_{err})} \tag{3.21}$$

lag_auto The lag-1 autocorrelation '*lag_auto*' is the dependence of the signal with itself shifted by one. It can be used to represent how similar consecutive points are.

$$lag\_auto = \frac{\sum_{i=2}^{N}(m_i - \bar{m})(m_{i-1} - \bar{m})}{\sum_{i=1}^{N}(m_i - \bar{m})^2} \tag{3.22}$$

AD The Anderson-Darling '*AD*' test is a statistical test for the similarity of a sample with a distribution (Anderson and Darling, 1952). Here, it is used to test for normality where $AD \to 0.25$ for a normal distribution.

std_nxs The normalised excess variance '*std_nxs*' Vaughan et al. (2003) is variability metric commonly used in Active Galactic Nuclei variability (Gliozzi et al., 2002; Vagnetti et al., 2016; Gonzalez et al., 2023).

$$std\_nxs = \frac{\sum_{i=1}^{N}(m_i - \bar{m})^2 - m_{err}^2}{N\bar{m}^2} \tag{3.23}$$

trans_flag The transient flag '*trans_flag*' is a boolean flag that is used to try to capture potential transients that are misidentified as periodic variable stars. The phase fold of a transient variable, such as a microlensing event, can look clean enough to be identified as periodic by both PDM and the neural network FAP. Figure 3.8 shows the raw and erroneously phase folded light curve for 'OGLE BLG-ECL-292071', which is misclassified as an eclipsing binary in Soszyński et al. (2016).

The transient flag is calculated at the time that the straight line is fitted to the cleaned light curve, figure 3.4. Each of the bins that is used for the straight line fit has the inter-quartile range (IQR) calculated. The median IQR for each of the bins is compared to each individual IQR for each bin. The transient flag is set to 1 if any bin has an IQR one third larger than the median IQR.

$$trans\_flag = \begin{cases} 1 & \text{if IQR}_i > 1.33 \times median(IQR) \\ 0 & \text{otherwise} \end{cases} \tag{3.24}$$

ls_bal_fap The Baluev FAP (Baluev, 2008) is a false alarm probability calculated from the analysis of the Lomb-Scargle periodogram. A discussion of the Baluev FAP can be found throughout chapter 2 where it is used as the primary comparison to the neural network FAP.

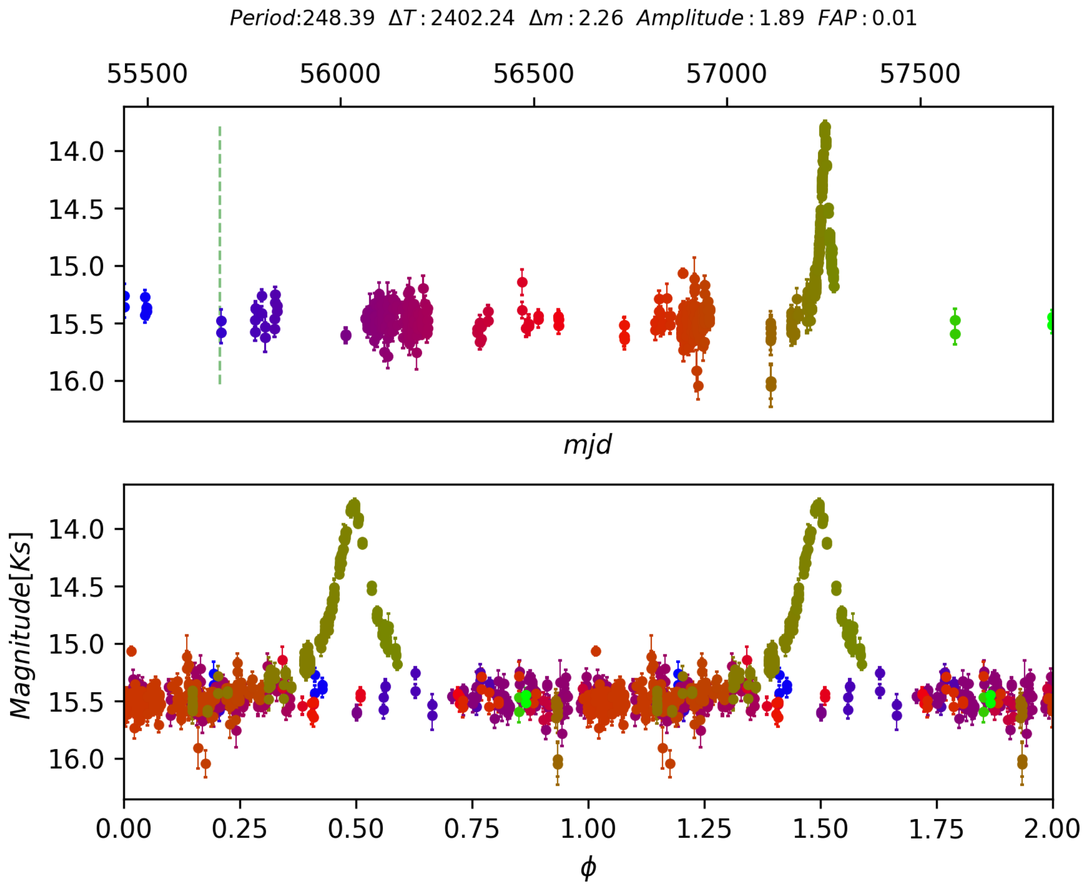


FIGURE 3.8: Top: Raw light cure showing the point of the transient event clearly at ≈ 57250 mjd. The green dashed line represents the period (i.e. $t_0 + Period$). Bottom: The incorrect phase fold at 248.39 days

Periodogram Statistics Each of the periodograms is analysed thoroughly for reliable peak extraction (section 3.5.2) The same analysis is performed for each of the LS, PDM and CE periodograms (the only difference being LS peaks are at a maximum value whereas both PDM and CE seek to minimise their value).

Each periodogram is analysed and the three most prominent peaks are extracted. From this process we have the period '$LS/PDM/CE\_period\_0,1,2$', peak value '$LS/PDM/CE\_y\_0,1,2$' (height of the peak) and peak width '$LS/PDM/CE\_peak\_width\_0,1,2$' for each of these three peaks. To allow for future comparison against the peak values, each periodogram has multiple percentiles calculated - $LS/CE/PDM\_0.001, 0.01, 1, 25, 50(median), 75, 99, 99.9, 99.99$

No such analysable periodogram is constructed for GPs and so we instead save all of the fitted metrics;

- 'gp_A' - amplitude factor of the covariance
- 'gp_l' - length scale of the RBF kernel

- ‘gp_g’ - ‘Γ’ the relative importance of the RBF kernel

- ‘gp_P’ - Period

3.7 The Catalogue

The PRIMVS catalogue has 86,507,172 computed sources at the time of writing. If we take a heuristic cut of anything with a FAP less than 0.3 we have 5,161,222 periodic variable stars. The true number of periodic variables is likely to be different than that. The FAP method discussed in chapter 2 was developed as part of this catalogue's pipeline. Hence, it has not been tested on large scale real data.

We can compare this to the VIVACE catalogue (Molnar et al., 2022). The VIVACE catalogue is a catalogue of periodic variables in VVV which this catalogue aims to supersede. Virtually all (97%) of the sources found in VIVACE can be found in this catalogue. Those that are not found in PRIMVS are because of different quality cuts

The University of Hertfordshire High Performance Cluster was used for the computation of each of the three tests. Each test was computed with 64 parallel instances with 4 cores and 2 GB of RAM (Totalling 256 cores and 128 GB of parallel computation use).

It is difficult to calculate a compute time for this catalogue as speed improvements and re-runs of tests create uncertainty. It takes an average of $0.9 - 1.1$ seconds per source to compute the whole pipeline inclusive of cleaning and post-processing statistics. This means it would take $\approx$ 16 days to process the catalogue with one test (ignoring cases where extra periodograms are computed).

The Bailey diagram (Bailey et al., 1919)– Logarithmic period versus amplitude–is a fundamental tool for characterising periodic variable stars. Figure 3.9 shows the Bailey diagram for all stars with FAP$<$ 0.3 in the PRIMVS catalogue.

The absence of stars at $\log_{10}(P) \approx 1.4$ is because we currently exclude periods on the diurnal and lunar time scale ($\approx$ 30 days). These will be re-added in the next version of PRIMVS but at the time of writing too many contaminants rendered this period range largely unusable. If we compare to the Bailey diagram constructed from Galactic bulge focused VIMOS data (Kains et al., 2019) we find similarly located densities. We see the same cluster of short period stars ($-1 < \log_{10}(P) < -0.5$) which are suspected contact binaries. We also see a density of stars where we expect to see Cepheids (da Silva et al., 2022; Kains et al., 2019; Bono et al., 2000). We do not see evidence for the typical 'double-peak' distribution caused by the Hertzsprung progression (Hertzsprung, 1926; Christy, 1975; Bono et al., 2000) (we should see a 'V' shape centred at $\sim$ 10 days). As we have not fully classified this catalogue it is likely that the lack of this shape is due to non-Cepheids, such as EBs and RR Lyrae, filling that gap.

We measure a completeness limit of 90% for our periodic variable stars to be at a magnitude of $\approx$ 14.5. Figure 3.10 shows the magnitude distribution for all stars in the PRIMVS catalogue with a FAP$<$ 0.3.

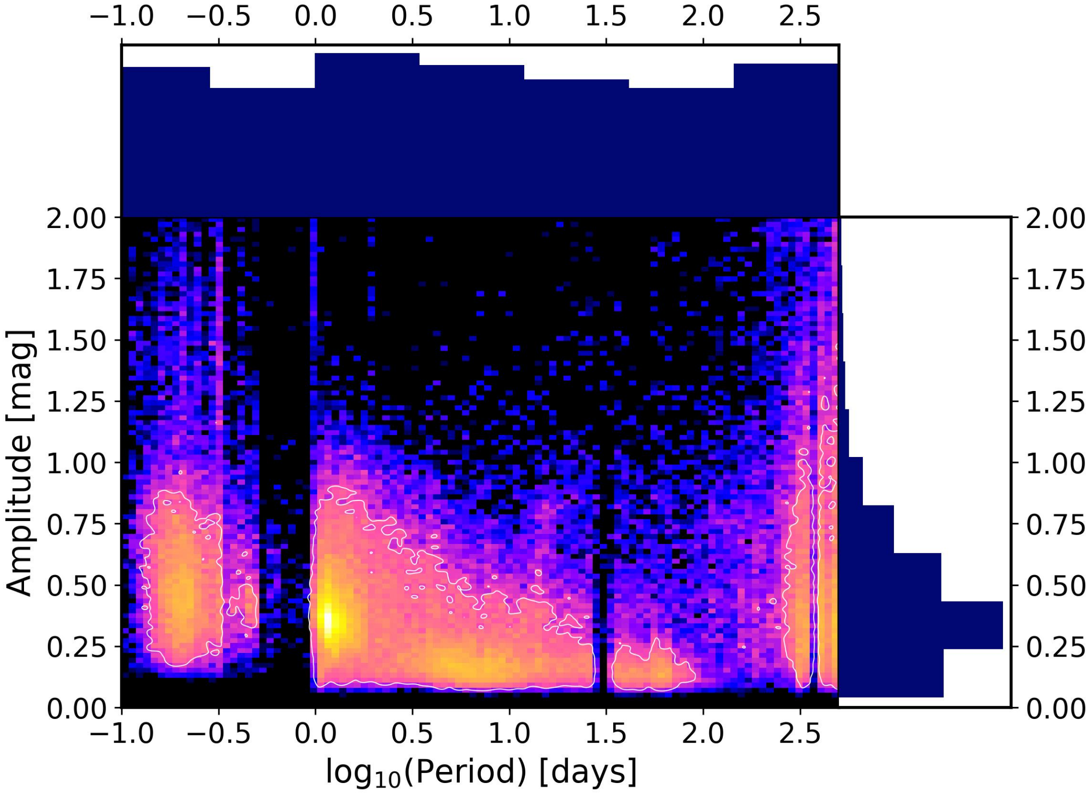


FIGURE 3.9: A plot of $\log_{10}(Period)$ verses Amplitude of all stars with an FAP< 0.3. The histograms for $\log_{10}(Period)$ and Amplitude are also displayed on their respective axes. For clarity with the large sample size, a 2D histogram is used with a contour around the 80th percentile of the data. The colour axis (show of density) of the 2D histogram is in log scale.

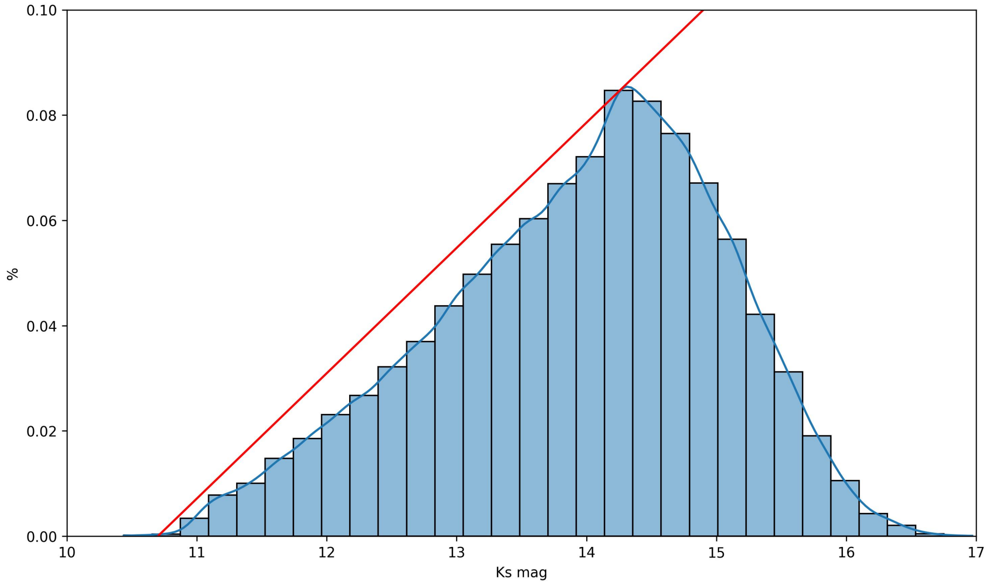


FIGURE 3.10: A histogram of magnitudes of all stars with a FAP< 0.3.

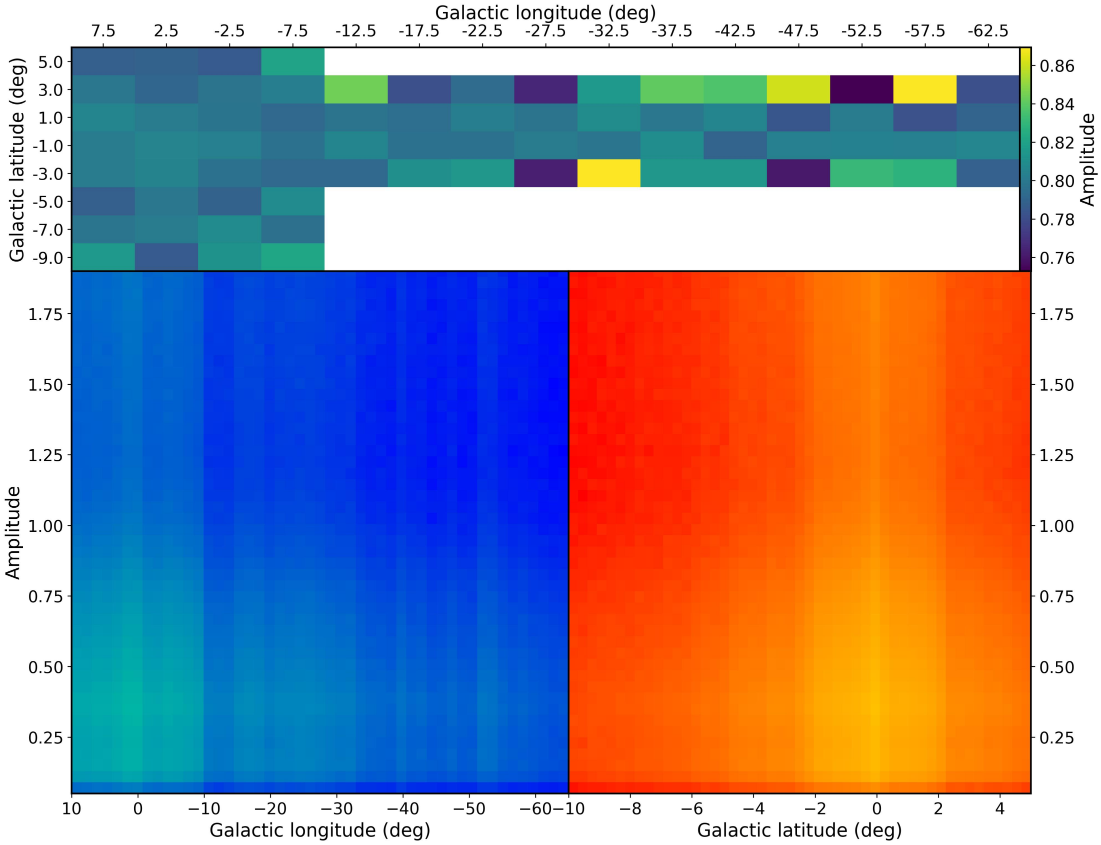


FIGURE 3.11: Light curve amplitude as a function of Galactic coordinates. Top: Histogram showing the median amplitude in each bin with respect to Galactic coordinates. Bottom Left: Density scatter plot showing amplitude as a function of Galactic longitude. Bottom Right: Density scatter plot showing amplitude as a function of Galactic latitude.

We describe the VVV survey as "*an infrared time-series survey focused on the southern viewable Galactic disk and bulge*" and so it is fitting we check the key parameters of the PRIMVS catalogue against their position in Galactic coordinates.

The bottom two panels of figure 3.11 show light curve amplitude as a function of Galactic latitude and longitude. Most objects are found to have an amplitude < 1. Figure 3.12 shows the magnitude distribution in the same way as figure 3.11. A homogeneous distribution can be seen throughout except for the Galactic bulge where the photometric depth increases by 0.2 mag. This same pattern of brighter sources can also be seen in figure 3.12. This is imparted from the array of detectors used in the VISTA telescope.

The observing pattern of the VVV survey reveals a correlation between position in Galactic coordinates and the specific detector used for measurement. This is apparent as the observing pattern for the VVV survey is based in Galactic coordinates. The VISTA telescope employs an array of 16 Raytheon VIRGO HgCdTe 0.84-2.5 micron detectors (Bornfreund, 2005), arranged as shown in Figure 1.8. As these detectors utilise relatively early-stage technology, they can exhibit differences in sensitivity, linearity, and particularly in saturation limits. This variability

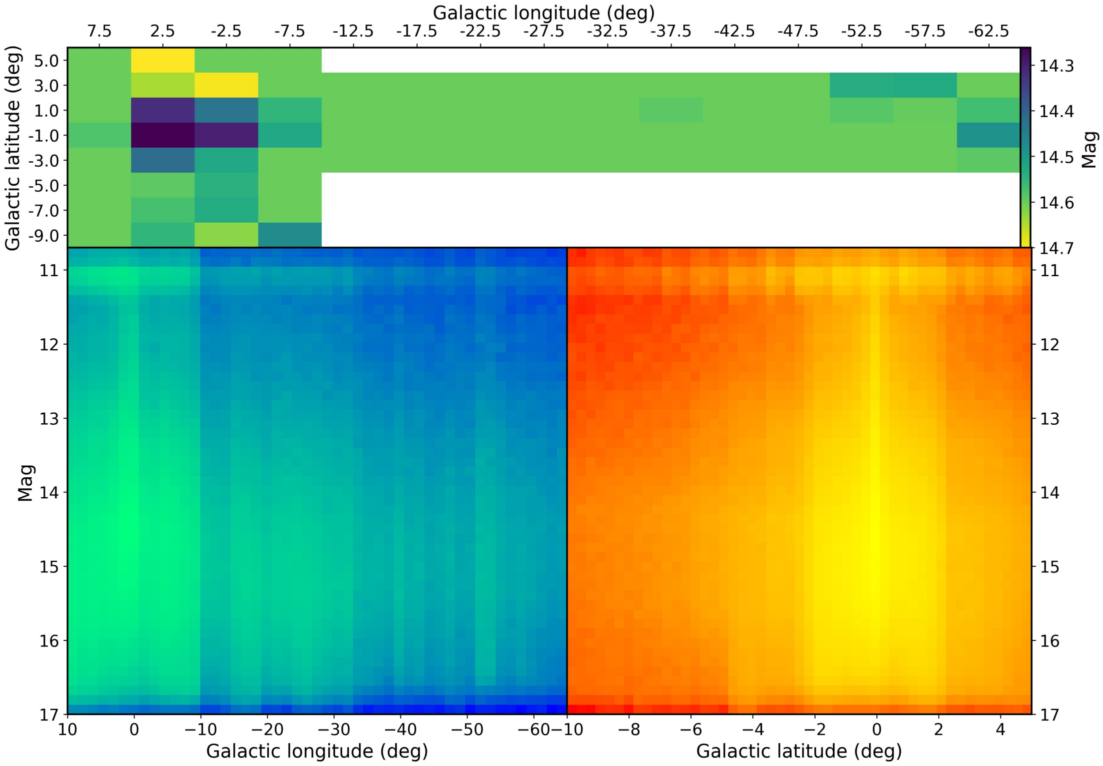


FIGURE 3.12: Light curve magnitude as a function of Galactic coordinates. Top: Histogram showing the median magnitude in each bin with respect to Galactic coordinates. Bottom Left: Density scatter plot showing magnitude as a function of Galactic longitude. Bottom Right: Density scatter plot showing magnitude as a function of Galactic latitude.

across the detectors affects the precision and reliability of measurements, as some detectors reach saturation at lower brightness levels than others. Such discrepancies can lead to areas of the VVV survey region which are probed to greater depth and/or brighter magnitude.

We can look for other features as a function of location that help us to begin to verify the completeness of PRIMVS. Due to the nature of period finding techniques, light curves with uneven magnitude distributions/non-sinusoidal shapes are often underrepresented in periodic variable catalogues. Eclipsing binaries are ubiquitous and largely homogeneous throughout the galaxy (Mowlavi et al., 2023). Due to the nature of an eclipsing binary light curve, they are largely unique in their light curve morphology and resulting magnitude distribution. This is exemplified by figure 3.13 where a typical EB ($\beta-$lyrae) light curve can be seen.

For an EB, a large distribution of the points are in the brighter stages of the light curve (either the lack of an eclipse or a relatively minor reflection) with fewer points tracing out the two eclipses. This results in a uneven distribution of points, unlike how a Cepheid, RR-Lyra or AGB light curve would have[4]. EBs are therefore likely to be the largest contributor to any measured skew deviating from 0 in the catalogue. Figure 3.14 shows the measured skew in PRIMVS as a function of Galactic coordinates. The median skew is greater than 0.2 across the whole of

[4] assuming no other perturbation

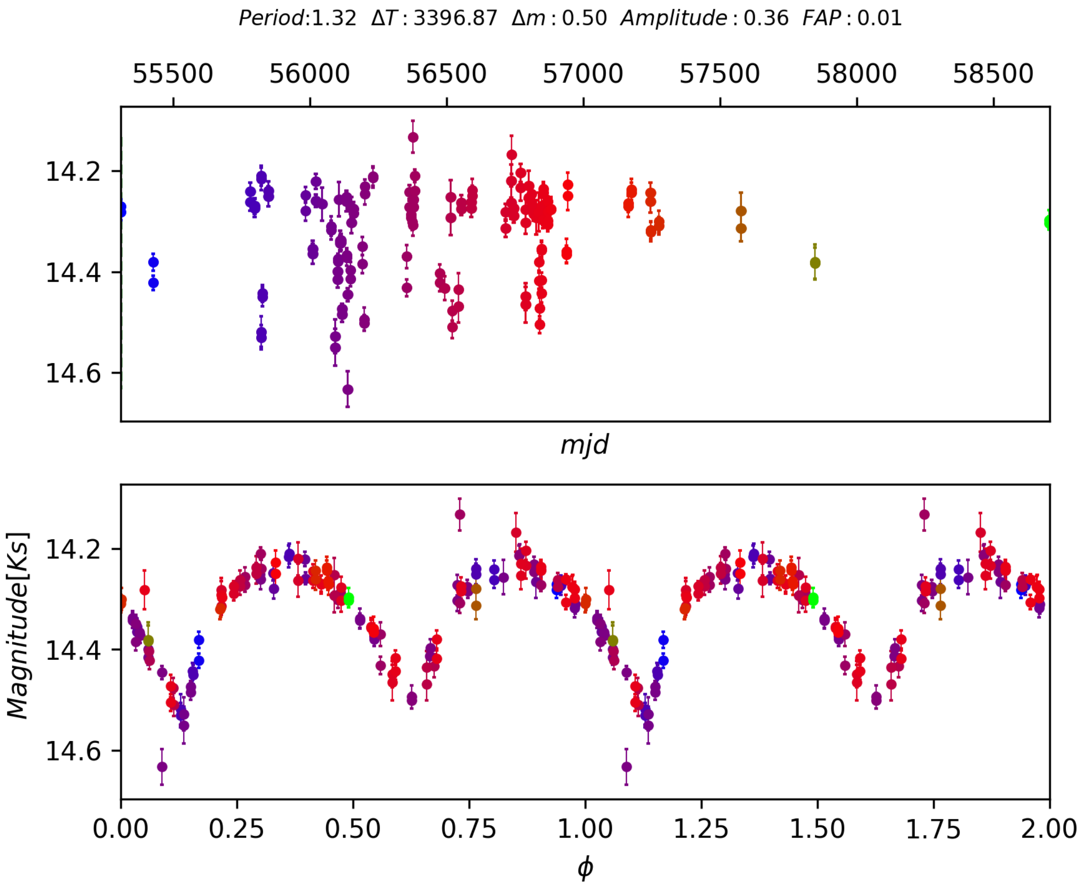


FIGURE 3.13: Light curve of eclipsing binary 'V* V2679 Sgr'. Top: Light curve as a function of time. Bottom: phase folded light curve. The colours are correlated with time.

the PRIMVS catalogue but also appears largely homogeneous throughout the Galactic disk and bulge. This helps to indicate that we have not preferentially selected EBs in either the Galactic disk or bulge, regions with different densities of stars.

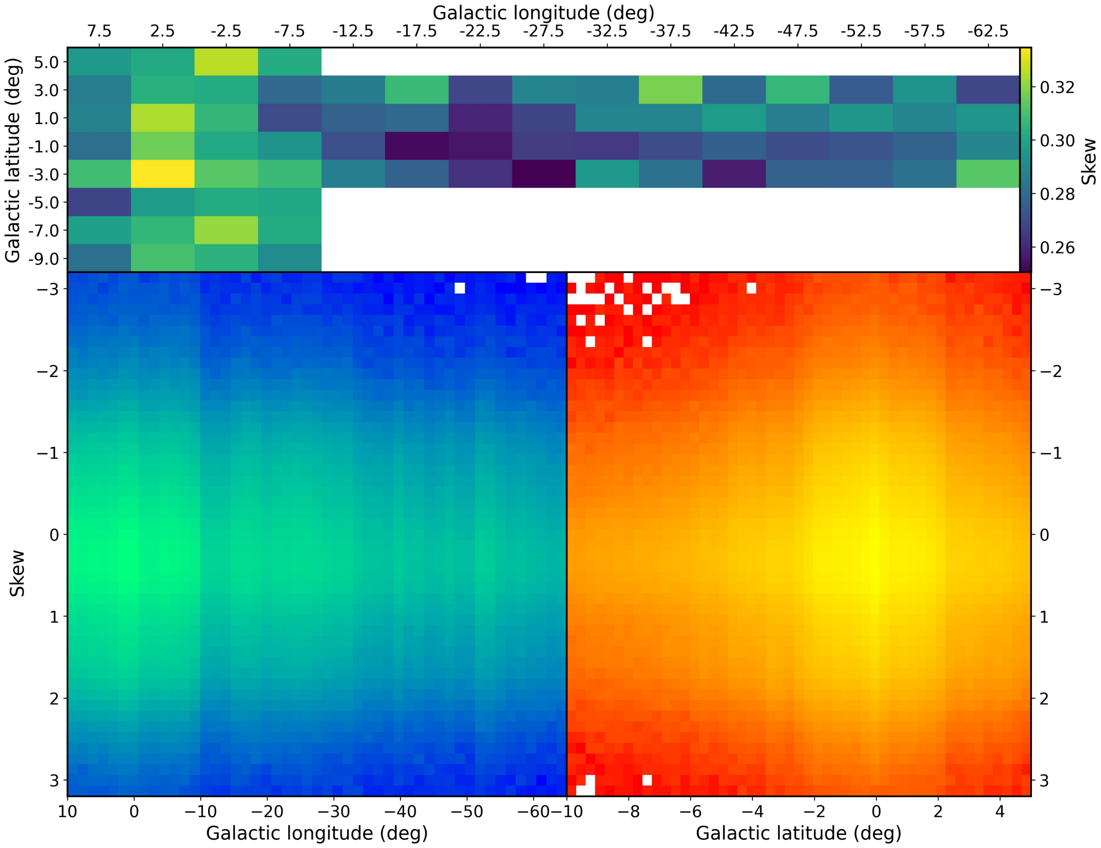


FIGURE 3.14: Light curve skew as a function of Galactic coordinates. Top: Histogram showing the median skew in each bin with respect to Galactic coordinates. Bottom Left: Density scatter plot showing skew as a function of Galactic longitude. Bottom Right: Density scatter plot showing skew as a function of Galactic latitude.

3.7.1 Quasi-periodic sources

The variability of a source can be from a multitude of non-mutually exclusive intrinsic and extrinsic reasons. It follows that there are many variable sources which feature some apparent quasi-periodicity. This quasi-periodicity could be from the combination of a periodic variability with some other aperiodic variability (as is commonly seen in YSOs) or single causes of quasi-periodicity (such as star spots). The treatment of quasi-periodicity was considered in the construction of this catalogue by the combination of feature that are outlined in section 3.6. The neural network FAP outlined in section 2 acts as something analogous to a measurement of structure in the phase folded light curve (similar to PDM). There exists cases where quasi-periodic sources feature strong structure within their phase fold. This will lead to the neural network based FAP to erroneously proscribe a FAP indicative of periodicity. However, other features that are calculated for the light curve will indicate a deviation from periodicity. Most notably is the neural network FAP and the Baluev FAP will likely disagree as Baluev FAP has a dependency on the similarity to a sinusoidal wave. It is likely that quasi-periodic sources live in a latent space described by each of these features that is exclusively different from the periodic sources. A future work for this project will be to identify this latent space region and assign a flag of quasi-periodicity to any sources that are within it.

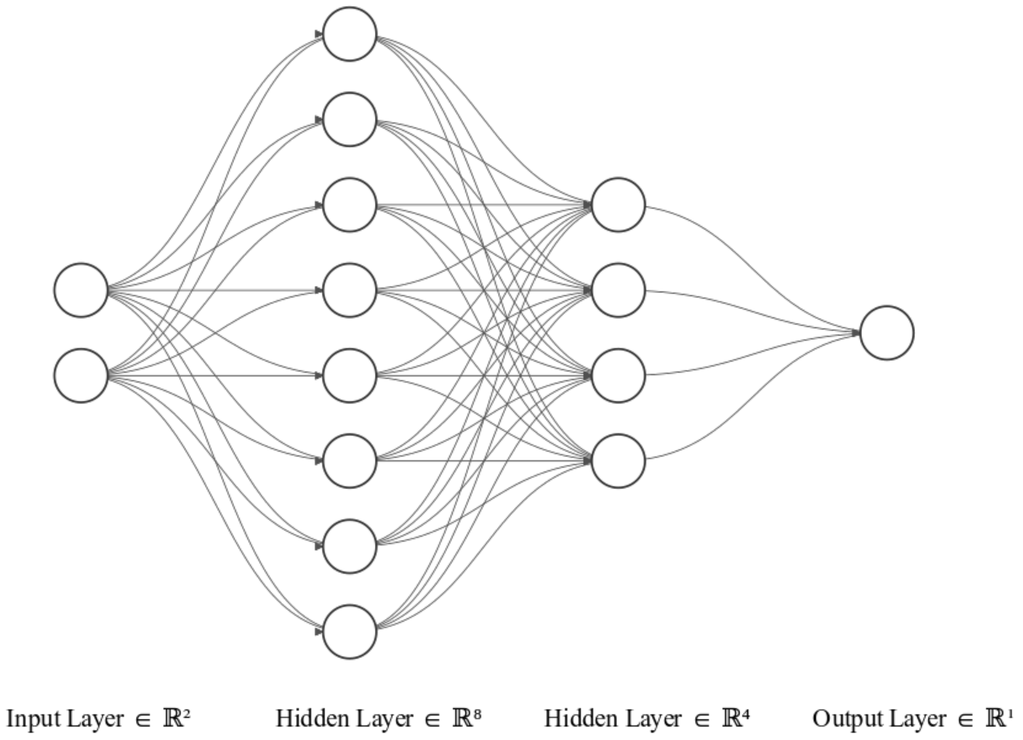

Input Layer $\in \mathbb{R}^2$
Hidden Layer $\in \mathbb{R}^8$
Hidden Layer $\in \mathbb{R}^4$
Output Layer $\in \mathbb{R}^1$

FIGURE 3.15: Showing architecture of the autoencoder that was used. For easy visualisation, each node here represents 16 actual nodes in the network.

3.7.2 PRIMVS Embedding

To continue with the overarching goal of unbiased exploration of the VVV data set we can make latent space representations of the PRIMVS catalogue to highlight potential groups. Figure 3.15 shows the architecture of the autoencoder (which takes the form of an MLP) that was used for this process.

The network was trained for 469 epochs with early stopping on the plateau of validation loss. The initial learn rate of 0.1 was halved each time the validation loss did not reduce for more than 10 steps. Only features with discernible physical meaning were used (i.e. features that an astronomer would use to begin to classify a star). Table 3.4 shows each of the features that were used. Features with a '*' are taken from the VIRAC catalogue. A future improvement for any methods using this selection of features would be to identify important and useless features. There is a lot of shared information between many of the features as a large portion of them are describing the magnitude distribution of the light curve.

Figure 3.16 shows the Uniform Manifold Approximation and Projection (UMAP; McInnes et al.,

*z med mag-ks med mag	Median z band magnitude - median ks mag
*y med mag-ks med mag	Median y band magnitude - median ks mag
*j med mag-ks med mag	Median j band magnitude - median ks mag
*h med mag-ks med mag	Median h band magnitude - median ks mag
*l	galactic longitude
*b	galactic latitude
Cody M	AM Cody 'M' value
stet k	Stetson 'K' value
eta e	Von Neumann 'eta e' vallue
med BRP	Median buffer range percentage
range cum sum	Range of a cumulative sum
max slope	Maximum slope between two points
MAD	Median Absolute Deviation
mean var	Mean Variance
percent amp	Percentage Amplitude
true amplitude	Amplitude
roms	RObust Median Statistic
p to p var	Peak-to-peak variability
lag auto	Lag-1 autocorrelation
AD	Anderson-Darling
std nxs	Normalized excess variance
weight mean	Weighted Mean
weight std	Weighted Standard deviation
weight skew	Weighted Skew
weight kurt	Weighted Kurtosis
mean	Mean
std	Standard deviation
skew	Skew
kurt	Kurtosis
true period	Period

TABLE 3.4: Table showing each of the features used in the embedding process. Features with a '*' are taken from the VIRAC catalogue.

2018) of these features. UMAP is a machine learning technique used for dimensionality reduction. It is particularly effective at preserving both the local and global structure of the data, making it useful for visualisation of high-dimensional datasets, similar to t-SNE (van der Maaten and Hinton, 2008). UMAP works by constructing a high-dimensional graph representing the data, then optimising a low-dimensional graph to be as structurally similar as possible, hence reducing dimensions while retaining the data's original structure.

Figure 3.17 shows the Principle Component Analysis (PCA) projection of these features. PCA is a statistical technique used for dimensionality reduction while preserving as much of the variance in the high-dimensional data as possible. It works by identifying the directions (principal components) along which the variance in the data is maximised. The data is $\mathbb{Z}$-normalised and the covariance matrix is across each feature is calculated. Eigenvectors of the covariance matrix are computed and sorted with respect to the magnitude of their eigenvalues, this forms the

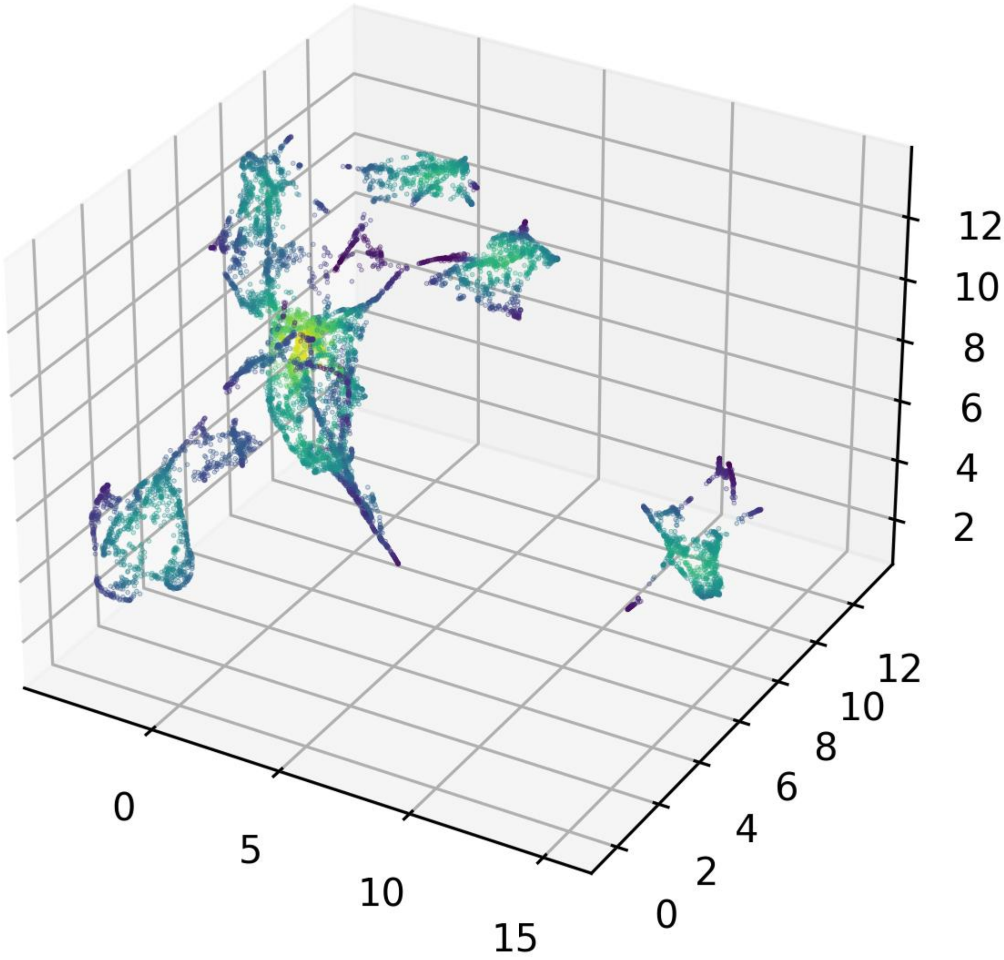


FIGURE 3.16: A 3 dimensional UMAP representation of the features from table 3.4

principle components. The original data is projected onto the principal components selected in the previous step, resulting in a new dataset with reduced dimensions. Figures 3.16 & 3.17 show 20,000 points with the colour representing density.

Both of these projections show similar features, most notably the smaller isolated group that can be seen in both. Comparing both projections shows that the same set of stars are found in both the PCA and UMAP isolated groups. These groups comprise $\approx 2\%$ of the total distribution in both projections. Comparing these groups to the rest of the distribution we find features which can parameterise it: amplitude > 1, Lag-1 autocorrelation < 0.2, $0.2 <$ FAP < 0.3 and, period $>$ 1000 days. Figure 3.18 shows the raw and phase folded light curves of 16 sources selected from this isolated group. These objects appear to be high amplitude, quasi-periodic variable stars. This is expected as the above parameterisation appropriately descibes such objects. It is also expected that the objects found in such an isolated group would not be as cleanly periodic.

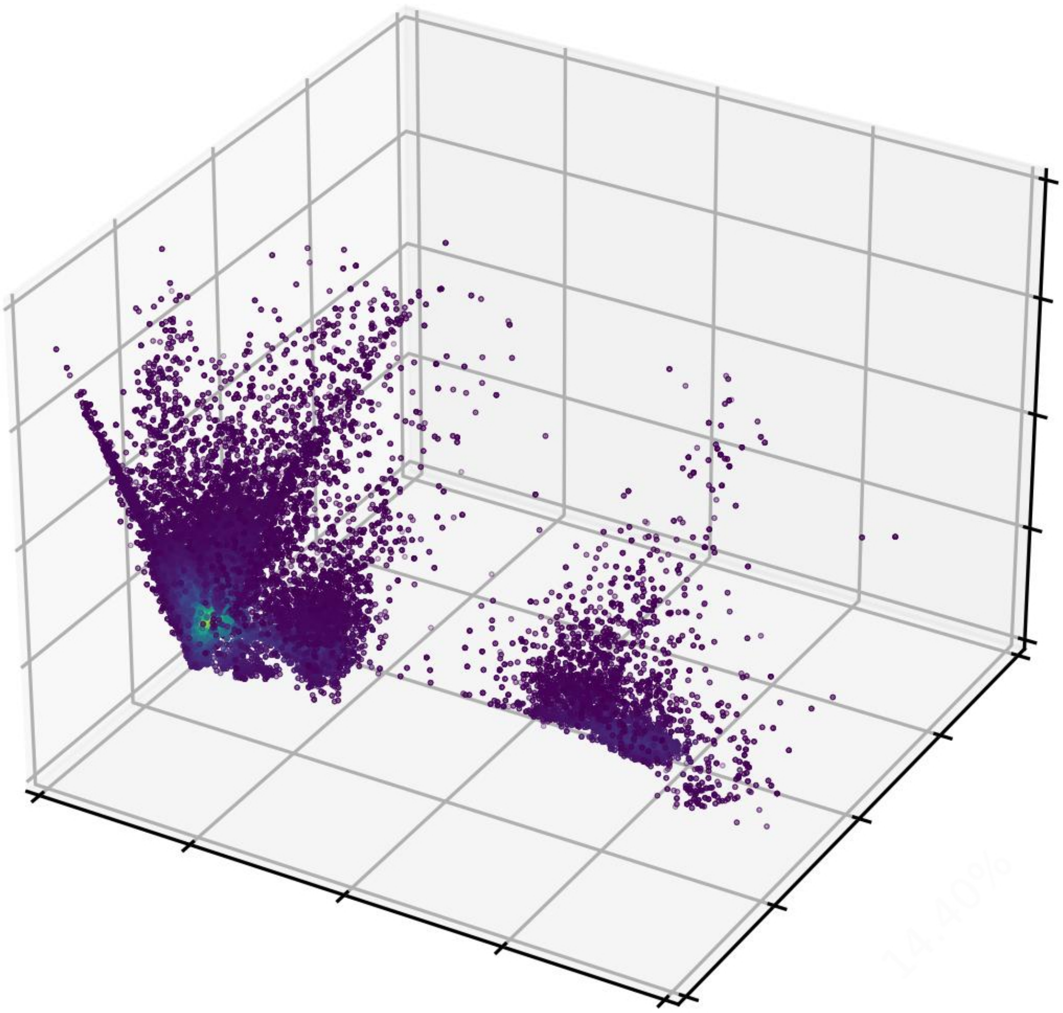
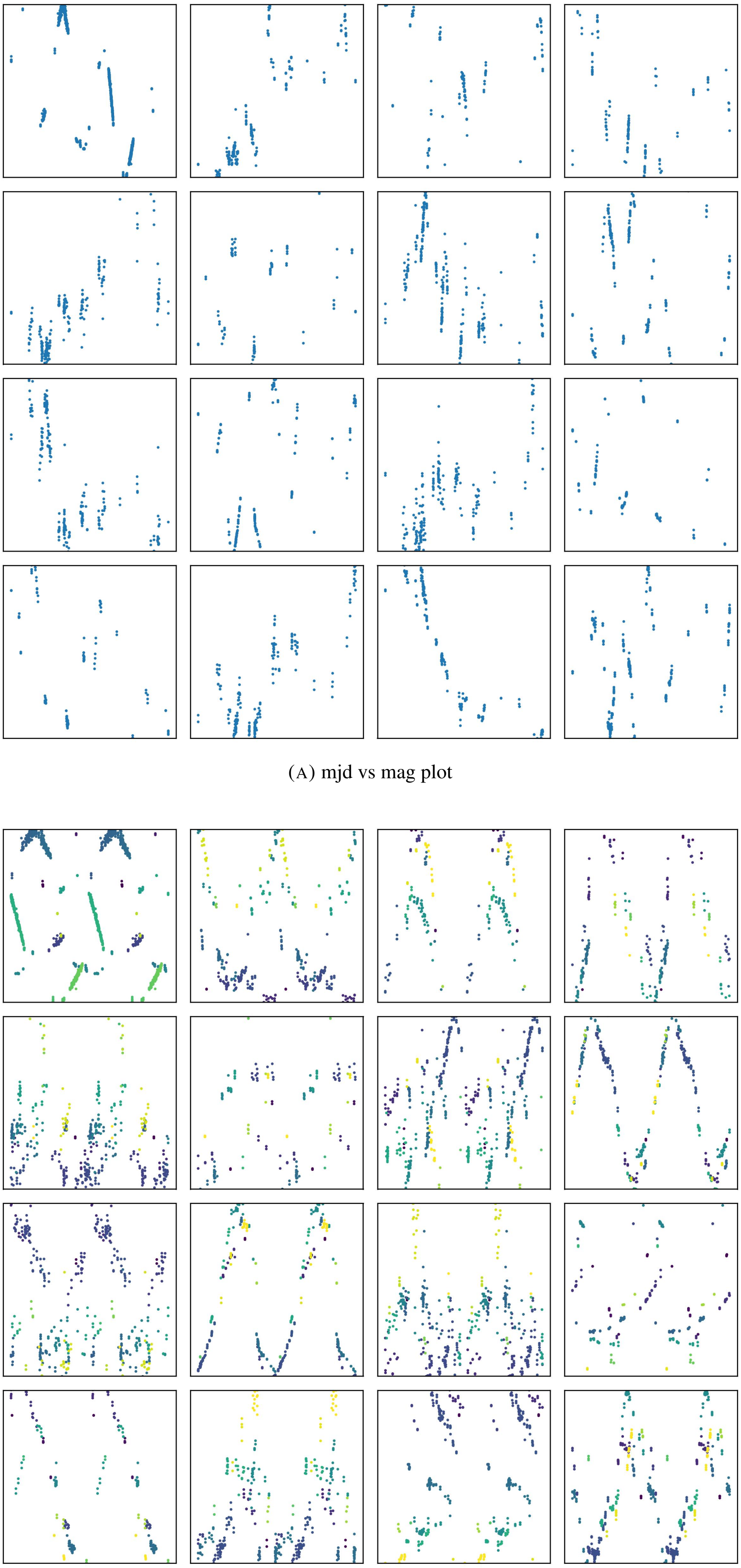


FIGURE 3.17: A 3 dimensional PCA representation of 31 features from table 3.4. The importance of the dimensions is; 32.58%, 14.40%, 13.68%

A Quasi-periodic (or Aperiodic) object would possess features more distinctly different than individual classes of periodic variable stars.

Interactive plots and data visualisations that are not suitable for the pdf format can be found at: https://nialljmiller.com/projects/PRIMVS/PRIMVS.html.



(A) mjd vs mag plot

(B) Phase vs mag plot (colour as mjd)

FIGURE 3.18: Light curves of 16 objects identified from the isolated group seen in figures 3.16 & 3.17

3.7.3 Decision Trees

The use of a pre-classified star catalogue as a training set for a machine learning algorithm is not new. The General Catalogue of Variable Stars (GCVS, (Samus' et al., 2017)) exemplifies such a catalogue, having compiled variable objects since 1946. This approach has been applied to data from VVV (Molnar et al., 2022), ASAS-SN (Jayasinghe et al., 2018, 2019), Gaia DR2 (Rimoldini et al., 2019) and DR3 (Rimoldini et al., 2023), EROS-II light curves (Kim et al., 2014), and in the identification of microlensing events (Husseiniova et al., 2021).

It follows that we can use this approach to classify the variable stars in the PRIMVS catalogue. This method is not without caveats however, most notably the biases inherited from the training set.

The cross match of PRIMVS with the Gaia DR3 all-sky classification catalogue (Rimoldini et al., 2023) yielded 118,172 sources with an FAP < 0.3 and a 'best class score' > 0.7. This selection forms the training set which was then used with a gradient boosted decision tree classifier 'XGBoost' (Chen and Guestrin, 2016). Figure 3.19 shows the distribution of classes found in the cross-matched data. This distribution is not even across all classes and reflects the selection bias of both the Gaia DR3 all-sky classification catalogue and the VVV PRIMVS catalogue. This is a notable caveat, especially as the differences in the data (e.g. optical vs near-IR) likely leads to different selection biases.

Figure 3.20 shows the confusion matrix achieved from using the Gaia DR3 all-sky classification catalogue to form our training set for classifying PRIMVS. It can be seen that the majority of classes are correctly identified with a high completeness. All of the White Dwarf and RCB variables are misclassified as EBs and LPVs respectively. RCB variables are hydrogen-poor, carbon/helium-rich, high-luminosity stars. Their variability is characterised by high amplitude (1-9mag) aperiodic changes on the order of hundreds of days. This is superimposed by periodic pulsations up to several tenths of a magnitude on the time scale of tens of days (Clayton, 1996). The light curves of RCBs are therefore complex and likely span a large range in any feature space. Considering this with their scarcity, it is not surprising we misclassify all of them as LPVs, a much more common class with similar features. Given this, RCBs and White Dwarfs were removed from the data.

We can calculate how confident the model is with the highest probable class. Figure 3.21 shows the 'Entropy' versus the 'Confidence metric' for each class with a probability> 0.5 for its most likely class. where 'Entropy' is the entropy across the classes, (i.e. $S = \sum P_{class} \ln(P_{class})$). Therefore, a lower Entropy suggests that the model's predictions are more certain because the probability distribution across classes is less uniform – One class has a much higher probability compared to others, indicating a strong preference by the model for that class. The 'Confidence metric' is the difference between the most likely and next most likely class.

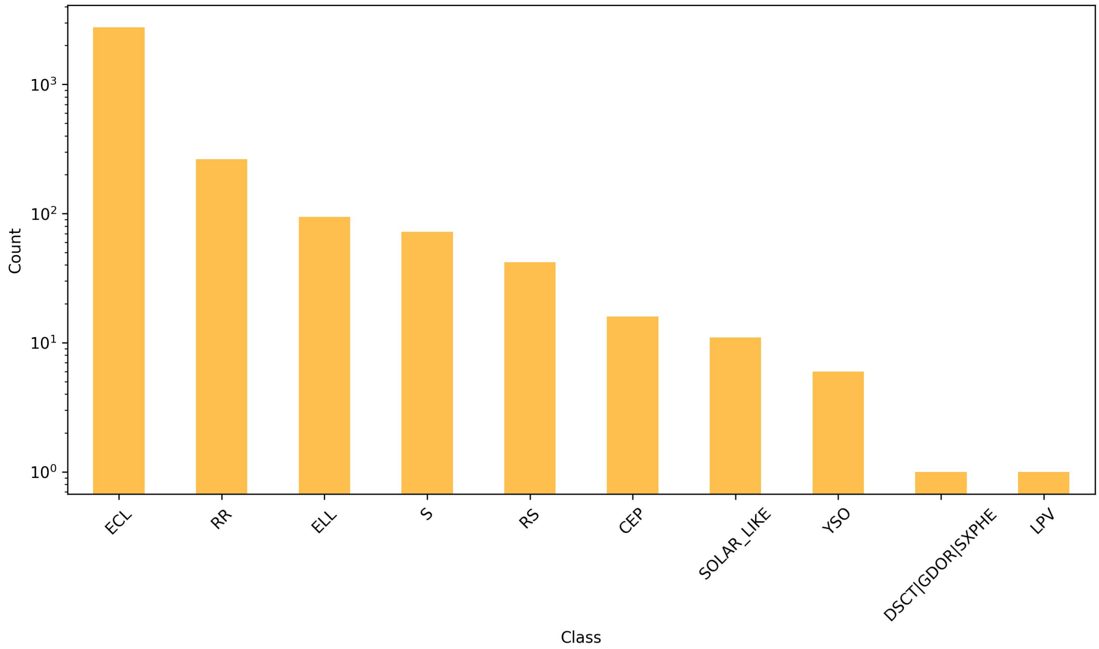
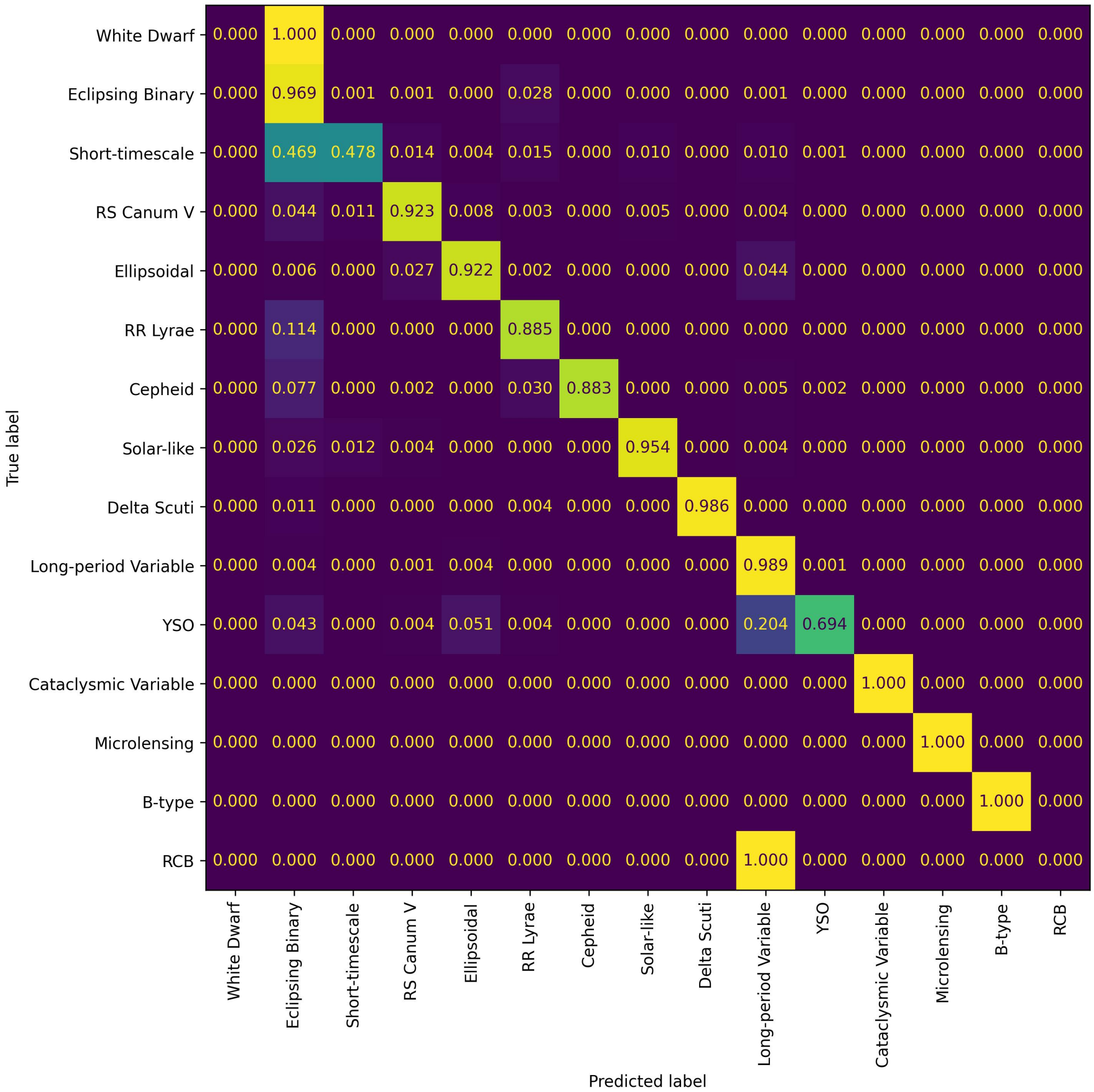


FIGURE 3.19: Training set of cross-matched VVV-Gaia data. Where; 'ECL' is eclipsing binaries, 'RR' is RR Lyraes, 'ELL' is Ellipsoidals, 'S' is short time-scale objects, 'RS' is RS Canum Venaticorum variables, 'CEP', is Cepheids, 'SOLAR_LIKE' is for Solar-like objects, 'YSO' is young stellar objects, 'DSCT||GDOR||SXPHE' is for delta-scuti like objects, and 'LPV' is for long period variables.

A Bailey diagram is an excellent tool for versifying how sensible our classifications are. Figure 3.22 shows the Bailey diagram constructed from the highest probability sources for each class. For both figure 3.22 and figure 3.23 we select only sources with a probability > 0.7, entropy < 0.2 and confidence metric > 0.9. The same colours and markers are also used to represent each class throughout figures 3.21,3.22, and 3.23 The Cepheid population is clearly visible and takes the expected form on the plot. The expected bimodal distribution of Cepheids can be seen as a loose 'V' shape at 10 days (Bono et al., 2000). We also see good agreement with Kains et al. (2019) in terms of our LPV, Delta Scuti and RR Lyrae placements.

Figure 3.23 shows the stellar classifications across the VVV survey region in relation to their positions within the Milky Way. This plot provides insights into the typical locations of different stellar populations. Cepheids, marked as red dots, are young, luminous stars commonly found in the thin disk throughout the galaxy Skowron et al. (2019). Figure 3.23 shows our sample of Cepheids throughout the disk mostly within $|l| < 1.5$, with an increased density at $|b| < 6$. Long-period Variables, such as Miras and semi-regulars are typically older, evolved stars and thus are more prevalent in the Galactic bulge and halo, where older stellar populations dominate (Wood and Bessell, 1983). Figure 3.23 shows these objects homogeneously spaced throughout the disk with significantly higher densities towards the inner bulge. RR Lyrae stars, yellow plus signs, are old, metal-poor stars found mainly in the Galactic bulge and halo, highlighting regions with



	White Dwarf	Eclipsing Binary	Short-timescale	RS Canum V	Ellipsoidal	RR Lyrae	Cepheid	Solar-like	Delta Scuti	Long-period Variable	YSO	Cataclysmic Variable	Microlensing	B-type	RCB
White Dwarf	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Eclipsing Binary	0.000	0.969	0.001	0.001	0.000	0.028	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000
Short-timescale	0.000	0.469	0.478	0.014	0.004	0.015	0.000	0.010	0.000	0.010	0.001	0.000	0.000	0.000	0.000
RS Canum V	0.000	0.044	0.011	0.923	0.008	0.003	0.000	0.005	0.000	0.004	0.000	0.000	0.000	0.000	0.000
Ellipsoidal	0.000	0.006	0.000	0.027	0.922	0.002	0.000	0.000	0.000	0.044	0.000	0.000	0.000	0.000	0.000
RR Lyrae	0.000	0.114	0.000	0.000	0.000	0.885	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cepheid	0.000	0.077	0.000	0.002	0.000	0.030	0.883	0.000	0.000	0.005	0.002	0.000	0.000	0.000	0.000
Solar-like	0.000	0.026	0.012	0.004	0.000	0.000	0.000	0.954	0.000	0.004	0.000	0.000	0.000	0.000	0.000
Delta Scuti	0.000	0.011	0.000	0.000	0.000	0.004	0.000	0.000	0.986	0.000	0.000	0.000	0.000	0.000	0.000
Long-period Variable	0.000	0.004	0.000	0.001	0.004	0.000	0.000	0.000	0.000	0.989	0.001	0.000	0.000	0.000	0.000
YSO	0.000	0.043	0.000	0.004	0.051	0.004	0.000	0.000	0.000	0.204	0.694	0.000	0.000	0.000	0.000
Cataclysmic Variable	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
Microlensing	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
B-type	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
RCB	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000

FIGURE 3.20: Confusion matrix of grouped classes from the Gaia Data Release 3 All-sky classification (Rimoldini et al., 2023). 'RCB' is R Coronae Borealis variables and YSO is Young Stellar Objects.

ancient star populations ,(Cabrera Garcia et al., 2023; Ramos et al., 2018). This seems to be in agreement with figure 3.23.

Similar to the autoencoder, interactive plots and data visualisations that are not suitable for the pdf format can be found at: https://nialljmiller.com/projects/PRIMVS/PRIMVS.html.

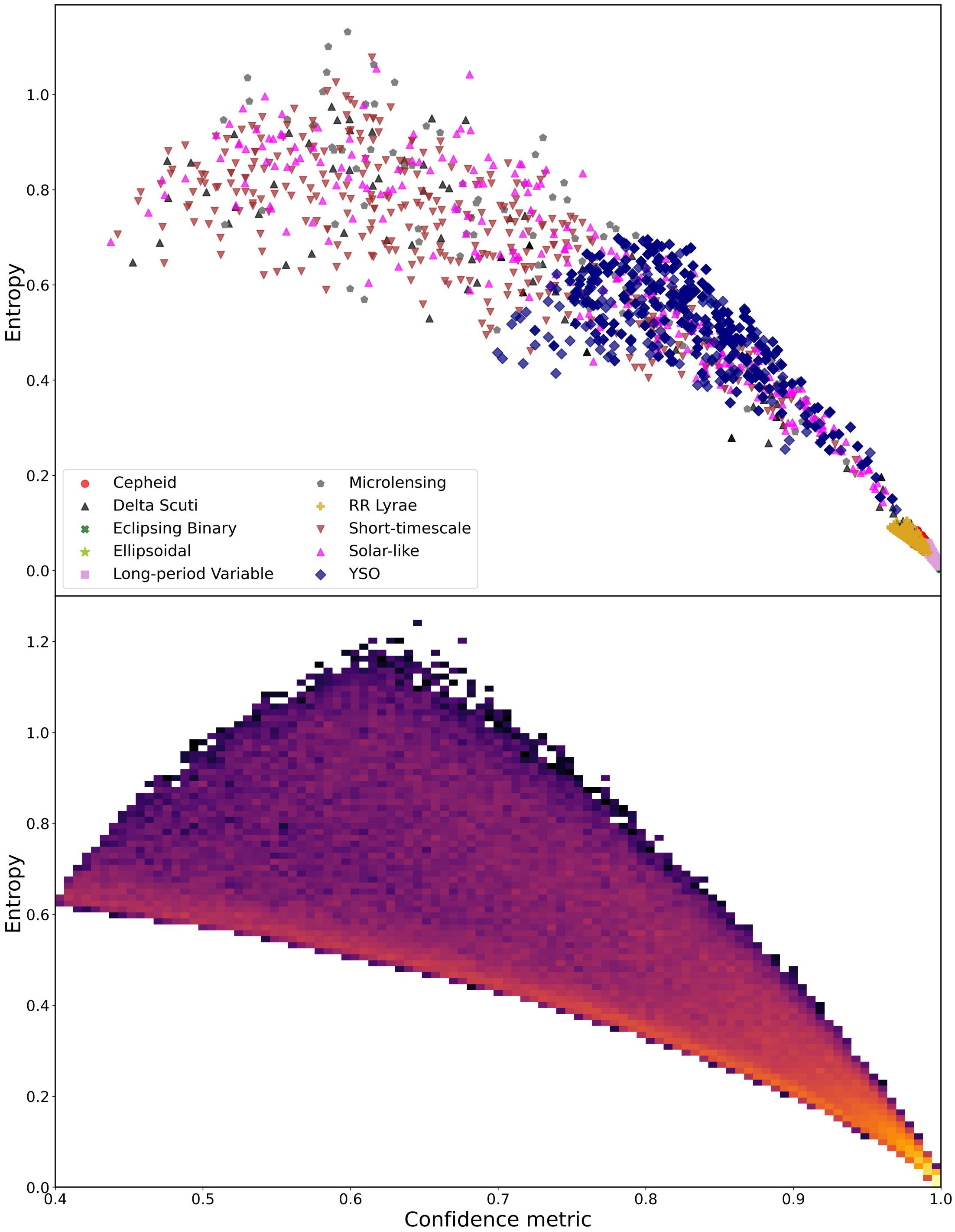


FIGURE 3.21: Top: A scatter plot of 'Entropy' versus the 'Confidence metric' for each each class with a probability>0.5. Bottom: A 2D histogram of 'Entropy' versus the 'Confidence metric' for the same data

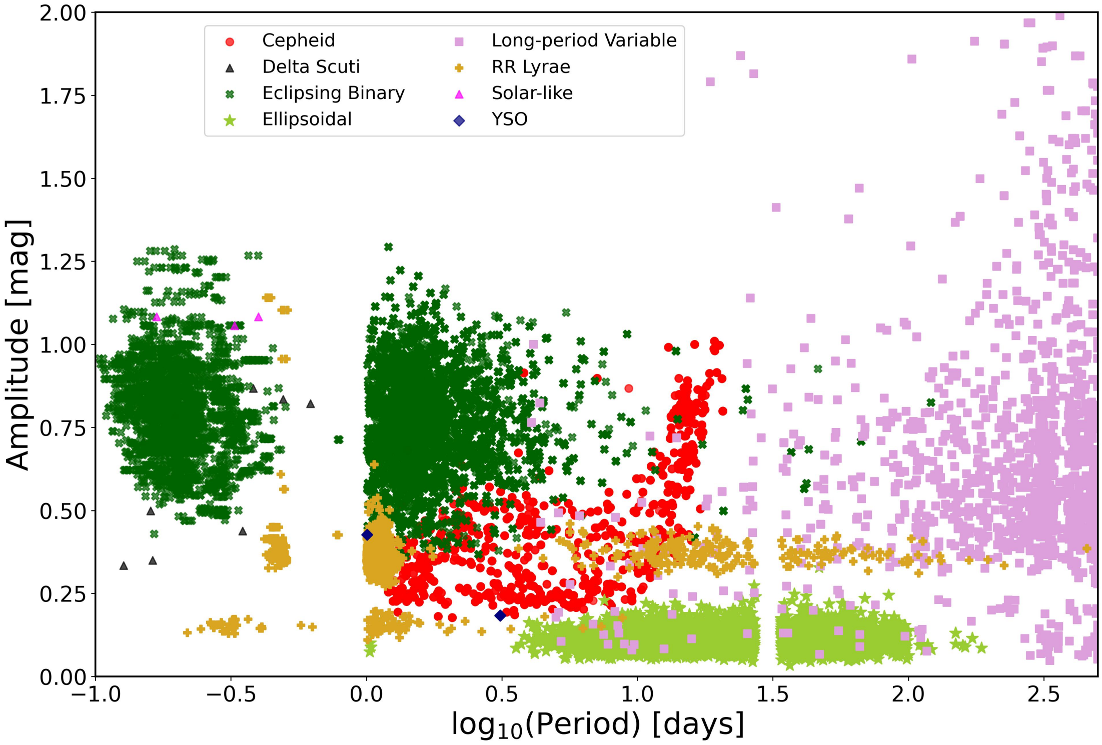


FIGURE 3.22: A plot of $\log_{10}(Period)$ verses Amplitude for the most confident predictions (top 10%) of each class from our decision tree which was trained using the Gaia DR3 all-sky classification catalogue.

3.8 Conclusions

This work introduces the PeRiodic Infrared Milky-way VVV Star-catalogue (PRIMVS), leveraging the VVV survey's depth and breadth to investigate the variability of astronomical sources within the Milky Way's Galactic bulge and disk. Through meticulous data cleaning and preprocessing, alongside modern analysis techniques, PRIMVS highlights the efforts towards an unbiased and complete identification and classification of periodic variable stars.

Our analysis employed various period-finding methods, demonstrating their strengths and weaknesses, and utilised a novel FAP method to enhance reliability in period identification. The catalogue includes over 86 million candidate variable sources and $\approx$ 5 million periodic variable stars.

Machine learning techniques, notably decision trees, have been shown as viable in classifying a substantial portion of PRIMVS sources. Cross-matched data from Gaia DR3 and the Simbad database has proven effective at identifying known and expected classes of stars. This approach, however, introduces its own set of challenges, notably the potential biases from the training sets and the limitation posed by Gaia's optical depth compared to the near-IR capabilities of VVV.

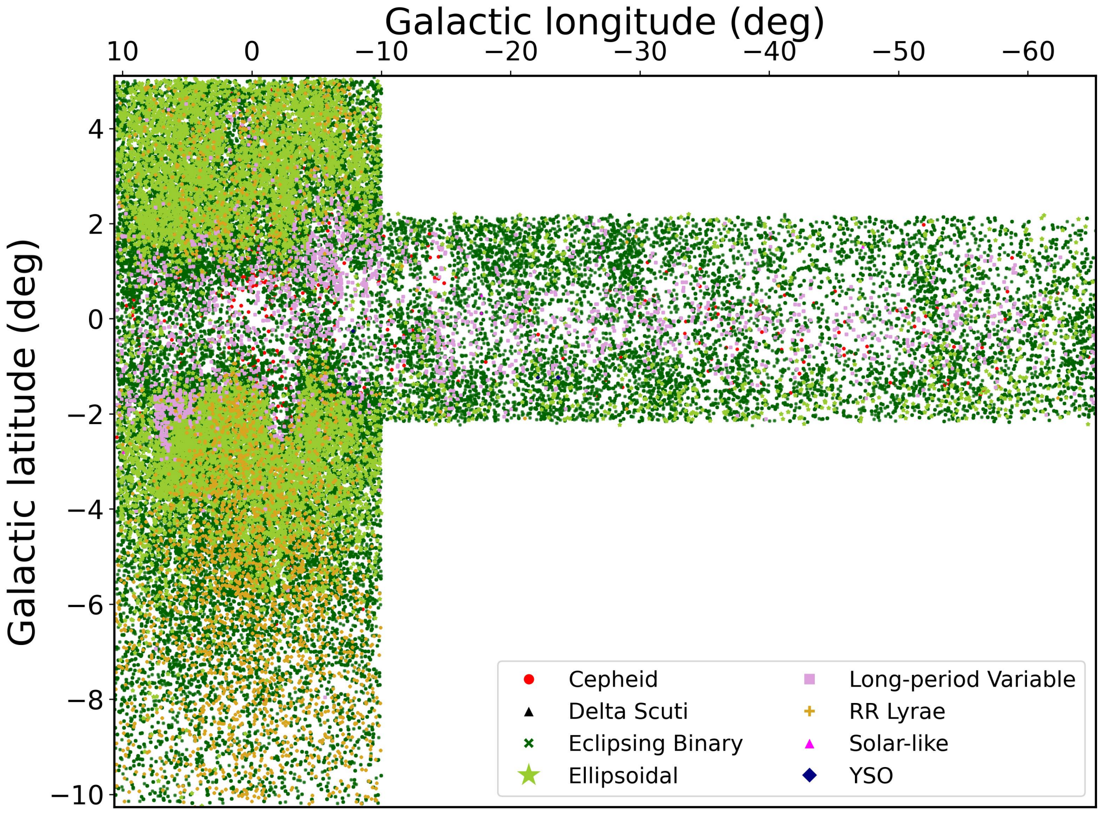


FIGURE 3.23: Spatial distribution of stellar classes across the VVV survey region in the context of the Milky Way. The decision tree based classification uses the Gaia DR3 all-sky classification catalogue as its training set.The absence of YSOs here is due to the conservative cuts shown in figure 3.21

PRIMVS not only advances our understanding of variable stars within the Milky Way but also showcases the potential of combining traditional astronomical analysis with modern data science techniques to explore and categorise astronomical sources effectively. Future work will aim to refine these classifications, expand the catalogue's scope, and further integrate deep learning approaches for a more thorough understanding of the stellar demographics and population of the Milky Way.

Chapter 4

Contrastive Curves

Abstract

We demonstrate that it is possible to extract semantically meaningful fixed length representations of stochastically sampled time series data. We use a novel neural network architecture (SimCLR with a gated recurrent neural network backbone) to go about this.

4.1 Introduction

The extremely large surveys typifying astronomy's Big Data Era will be impossible to parse manually (Minniti et al., 2010; Ivezić et al., 2019; Dewdney et al., 2009). If we are to consistently interrogate this data deluge at scale we need to devise reliable and robust automated methods. Deep learning has already gained a foothold in many data intensive fields, from astronomy, to particle physics, to chemistry. Deep learning is therefore a natural solution to astronomy's inherent scaling problem.

While supervised deep learning has been applied again (Storrie-Lombardi et al., 1992), again (Belokurov et al., 2003), and again (Charnock and Moss, 2017) in the quest to classify astronomical objects, its uses are limited by the availability of high quality labelled data. If there is no reliably labelled dataset one must turn to unsupervised or self-supervised methods to sort known categories of objects, and also to the find the 'unknown unknowns'—objects so obscure that they defy classification.

Self-supervised representation learning has recently exploded in popularity, with a slew of models being developed in rapid succession (i.e. Chen et al., 2020; Chen et al., 2020a; Grill et al., 2020; He et al., 2019; Chen et al., 2020b). At its core, representation learning attempts to produce semantically meaningful compressed representations (or embeddings) of complex highly dimensional data. Aside from simply being a compression device, these embeddings can also be taken and used in downstream tasks, like clustering, anomaly detection, or classification.

In recent years, pioneering work has applied self-supervised contrastive learning models to galaxy image clustering. Abul-Hayat et al. (2020) trained a simple framework for contrastive learning representations,(SimCLR; Chen et al., 2020) on multi-band galaxy photometry from the Sloan Digital Sky Survey,(SDSS; York et al., 2000). They demonstrated that the resulting embeddings capture useful information by using them directly in a training set for a galaxy morphology classification model and a redshift estimation model. Similarly, Sarmiento et al. (2021) trained a SimCLR model on integral field spectroscopy data from galaxies in the Mapping Nearby Galaxies at Apache Point Observatory survey (MaNGA; Bundy et al., 2015). They also found that SimCLR produces semantically meaningful embeddings. With these recent successes in mind, we ask: can we also use contrastive learning to interrogate astronomical time series data? In this work we address this question and leverage self supervised contrastive learning to explore the VISTA Variables in the Vía Láctea survey (VVV; Minniti et al., 2010).

In a concurrent work Donoso-Oliva et al. (2022) approach the problem of time series representation learning from a natural language processing (NLP) perspective. They repurpose the BERT (Bidirectional Encoder Representations from Transformers) Transformer network, which was initially developed in the context of NLP (Vaswani et al., 2017; Devlin et al., 2019). They then perform a 'pretraining' task on light curves, using the network to fill in zeroed datapoints within the time series. Once this pretraining task is completed, semantically meaningful embeddings can be extracted from the transformer network. Donoso-Oliva et al. (2022) show that these embeddings are useful for the downstream task of classification.

4.2 Contrastive self-supervised learning

Figure 4.1 describes a simple contrastive learning model in the vein of SimCLR (Chen et al., 2020) (this will be referred to as 'contrastive curves' throughout). This model takes as input a sample ($\mathbf{x}$) from the training set, and augments it to produce $\mathscr{A}(\mathbf{x})$. This augmentation is performed in such a way that $\mathscr{A}(\mathbf{x})$ shares enough semantically meaningful data with $\mathbf{x}$ to belong to the same class of objects. In the contrastive learning literature $(\mathbf{x}, \mathscr{A}(\mathbf{x}))$ is known as a positive pair. This positive pair is then passed to a Siamese neural network Φ, which projects the high dimensional input data onto a lower dimensional latent space. All other training set samples are assumed to belong to a different class to $\mathbf{x}$, and so can be combined with $\mathbf{x}$ to produce 'negative pairs'.

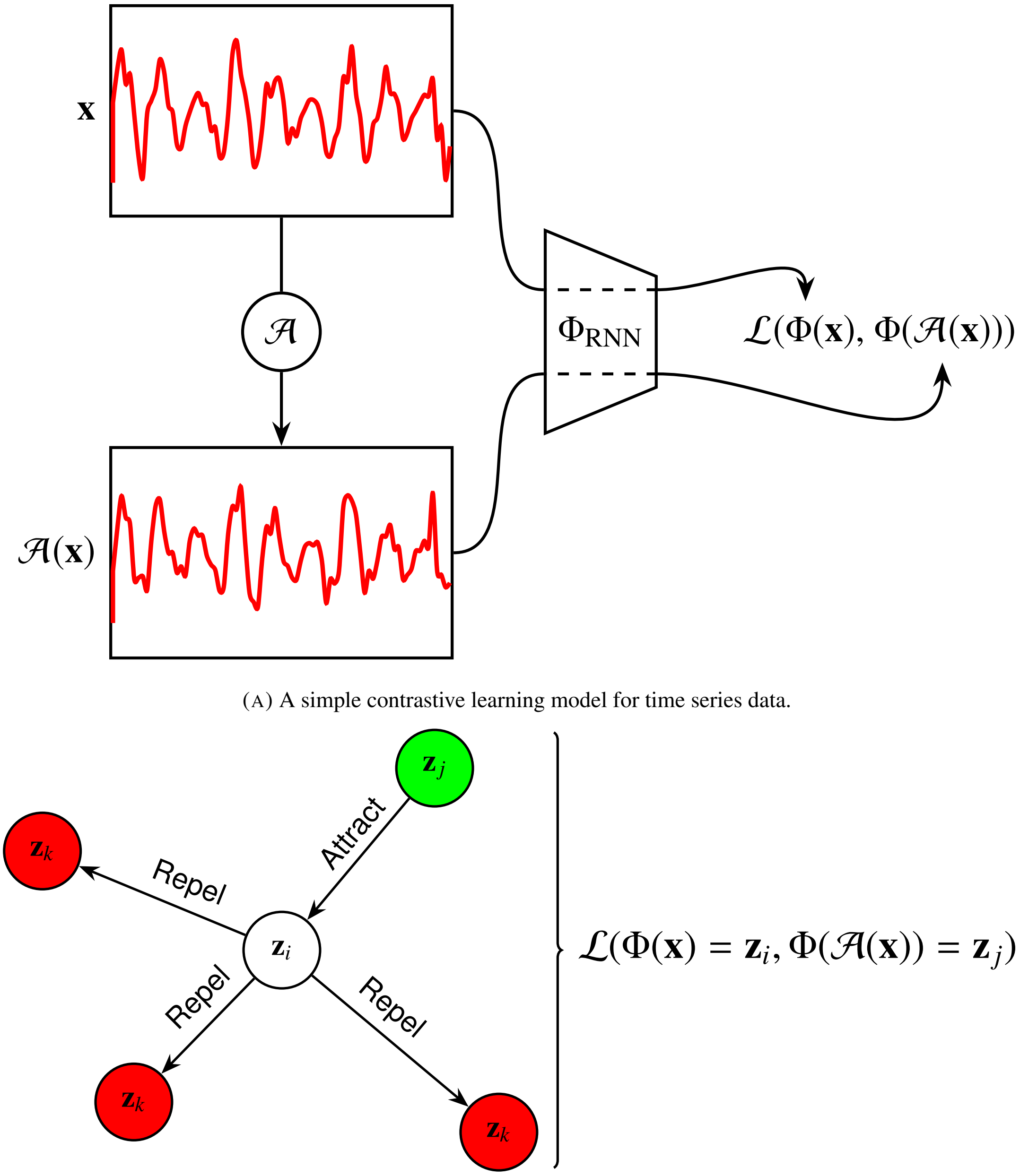
We use the normalised temperature cross entropy (NT-Xent) loss as our contrastive loss. The NT-Xent loss was first introduced in Sohn (2016), and was subsequently popularised by Chen et al. (2020). The NT-Xent loss is defined as

$$\mathscr{L}(\mathbf{z}_i, \mathbf{z}_j) = -\log\left(\frac{\exp(\mathbf{z}_i^T \mathbf{z}_j/\mathscr{T})}{\sum_{k=1}^{2N}(1-\delta_{ki})\exp(\mathbf{z}_i^T \mathbf{z}_k/\mathscr{T})}\right), \tag{4.1}$$

where $\mathbf{z}_i$ and $\mathbf{z}_j$ are a positive pair, and $\mathbf{z}_i$ and $\mathbf{z}_k$ are a negative pair. All embeddings are normalised. $\mathscr{T}$ is a 'temperature' hyperparameter introduced in Chen et al. (2020) to help the model learn from hard negatives. δ is the Kronecker delta.

As shown in figure 4.1b, minimising the NT-Xent loss minimises the distance in the embedding space between positive pairs while simultaneously maximising the distance between negative pairs. Therefore, once training is completed we expect to have moulded a semantically meaningful embedding space with similar vectors clustered close together.

In figure 4.2, we show a representation of our chosen model: a stacked bidirectional gated recurrent unit (GRU),(Cho et al., 2014). Due to the variable lengths of our input time series, we use a recurrent neural network. By taking the hidden states of our neural network, we convert the variably lengthed light curves to a fixed-length representation. Once we have this fixed-length representation, we can follow Chen et al. (2020) and use a single hidden layer fully connected neural network (i.e., $\mathbf{z} = g(\mathbf{h}) = W_2, \texttt{ReLU}(W_1\mathbf{h})$) to project the representation onto a final 64-dimensional space. We train on the vectors in this final space.


$\mathbf{x}$
$\mathcal{A}$
$\mathcal{A}(\mathbf{x})$
Φ_{RNN}
$\mathcal{L}(\Phi(\mathbf{x}), \Phi(\mathcal{A}(\mathbf{x})))$
$\mathbf{z}_j$
Attract
$\mathbf{z}_k$
Repel
$\mathbf{z}_i$
Repel
Repel
$\mathbf{z}_k$
$\mathbf{z}_k$
$\mathcal{L}(\Phi(\mathbf{x}) = \mathbf{z}_i, \Phi(\mathcal{A}(\mathbf{x})) = \mathbf{z}_j)$

(A) A simple contrastive learning model for time series data.

(B) The NT-Xent loss incentivises attraction in the latent space between similar examples while simultaneously incentivising repulsion between dissimilar examples.

FIGURE 4.1: In figure 4.1a a simple contrastive learning model is applied to time-series data. $\mathscr{A}$ is an augmentation pipeline. $\mathscr{A}$ could consist of noise addition, stochastic temporal shifting, and random data deletion. Φ is a function approximator that projects inputs onto an embedding space. Φ is typically a neural network; when processing time-series data Φ could be a recurrent neural network (RNN; McCulloch and Pitts, 1943). The loss $\mathscr{L}$ measures the distance between the embeddings $\Phi(\mathbf{x}) = \mathbf{z}_i$ and $\Phi(\mathscr{A}(\mathbf{x})) = \mathbf{z}_j$, and we train by attempting to minimise this distance while maximising the distance between dissimilar samples (figure 4.1b).

Our model is written in PyTorch (Paszke et al., 2019) and is available under the GNU Affero General Public License v3.0 at `https://github.com/nialljmiller/contrastive-curves`.

In Deb and Singh (2009) a different method is used to create representations of light curves without any priors. Their work presents a methodology for analysing light curves of variable stars, with both Fourier decomposition and PCA as principal analytical tools. This methodology is particularly designed to address the challenges posed by the non-uniform sampling of light curves, which is a common issue in observational astronomy. The authors implement a preprocessing step that involves phase-folding the light curves based on the stars' periods, then interpolating the magnitudes to achieve uniform sampling across the phase from 0 to 1 in steps of 0.01. This process ensures that each light curve is represented by a uniformly spaced set of points, making the data compatible with Fourier decomposition, which requires uniform sampling. For the Fourier decomposition analysis, the method transforms the light curves into a sum of cosine and sine series, thus allowing for the characterisation of the light curve through the Fourier parameters. This method has limitations however, Fourier analysis is fundamentally a fitting technique. While we can increase the Fourier terms to ultimately approximate any shape of light curve, this technique is computationally expensive and difficult to tune (BAART, 1982). PCA is employed as a more scaleable solution for analysing and classifying variable stars within large datasets. By directly using the interpolated light curve magnitudes as input, PCA bypasses the need for pre-computing Fourier coefficients, offering a significant advantage in terms of computational efficiency. The PCA transforms the original dataset into a new set of uncorrelated variables (principal components), which represent the most significant patterns within the data. However, the PCA method relies on linear assumptions about the data it analyses, which might not always be suitable for variable star light curves where non-linear phenomena govern brightness variations. Contrastive curves utilises data augmentation techniques to create positive pairs from the original data. This approach is particularly adept at capturing nuanced similarities between light curves that PCA might overlook due to its linear transformation and variance-focused dimensionality reduction. In contrast, contrastive curves, especially when implemented with neural networks such as bi-directional gated recurrent units (GRUs), can capture complex non-linear relationships within the data. This capability allows for a more nuanced understanding of the underlying astrophysical processes reflected in the light curves. Contrastive curves aims to create a semantically meaningful embedding space where similar examples are clustered together while dissimilar examples are repelled from each other. This approach can be particularly advantageous for classifying variable stars into their correct classes based on the intrinsic properties of their light curves. PCA, on the other hand, might not yield an embedding space that is as intuitively interpretable in terms of semantic similarity, as it primarily focuses on variance maximization.

To train our contrastive curves model, we generate slightly altered versions of each light curve. These alterations, or augmentations, include adding noise, randomly shifting time points, and

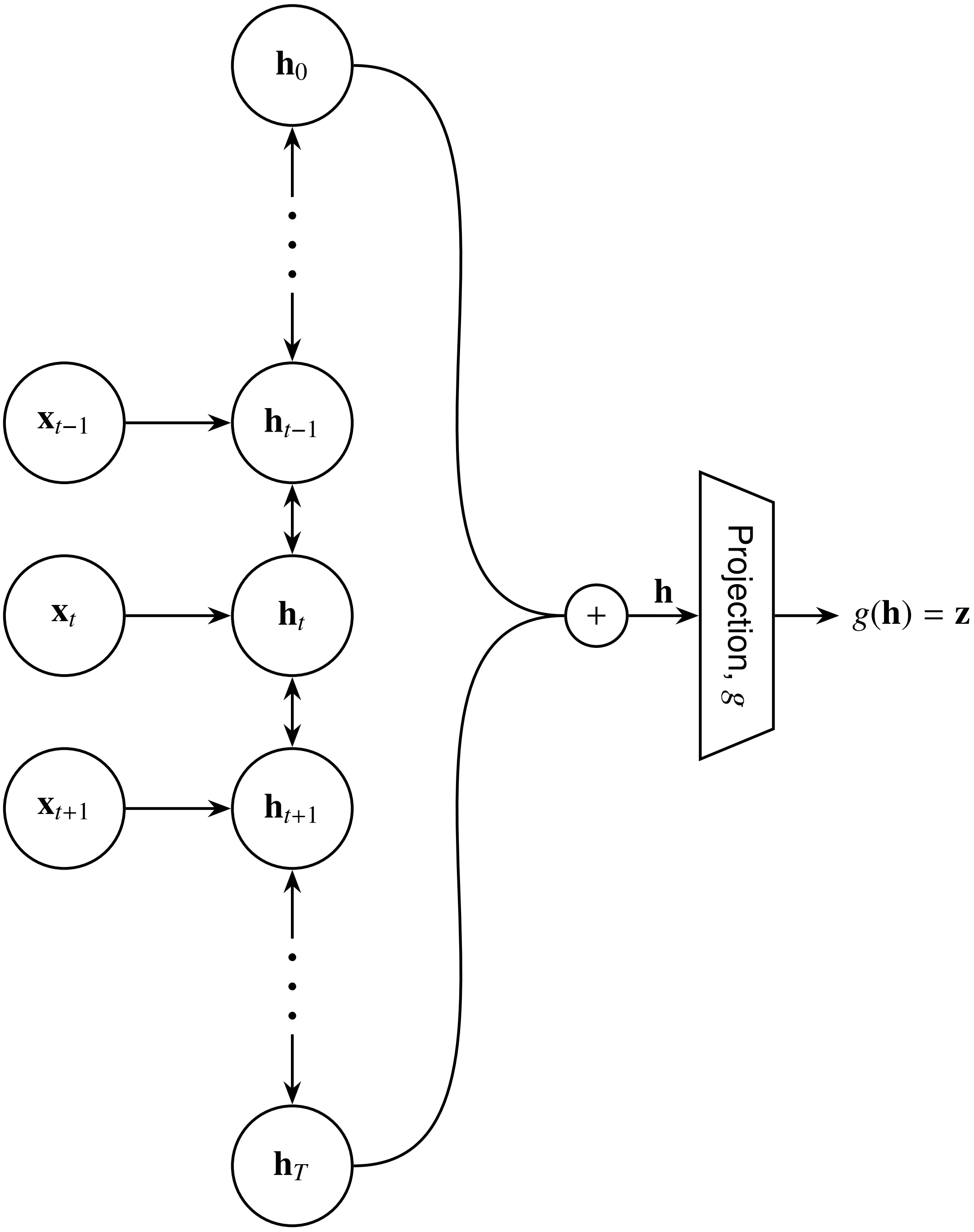


FIGURE 4.2: A variable star is input into our model. The star's time series is denoted $\mathbf{x}_t$. The time series is first passed into a bidirectional GRU network with an initial hidden state denoted $\mathbf{h}_0$ and a final hidden state denoted $\mathbf{h}_T$. The initial and final hidden states are concatenated along the channel axis, and the resulting vector $\mathbf{h}$ is passed through a linear projector. The output vector $\mathbf{z}$ is used for training (Eqn. 4.1). At inference time we follow Chen et al. (2020) and take $\mathbf{h}$ as the representation.

cropping sections. These augmentations are chosen to represent changes to a light curve that we could see between light curves of identical classes of variable stars. The purpose of these augmentations is to teach the model to recognise the essential features of the light curves despite these changes. Each original light curve and its augmented version form a positive pair, meaning they should be recognized as similar by the neural network. Conversely, each original light curve paired with different light curves forms negative pairs, which the model should recognise as dissimilar. The GRU, which comprises the majority of the network architecture, processes the input light curves and transforms them into fixed-length vectors, or embeddings. These embeddings are lower-dimensional representations that capture the significant features of the light curves.

The NT-Xent loss function helps the network learn by making sure that the embeddings of positive pairs are close together, while the embeddings of negative pairs are far apart. After training, the model can take any new light curve and convert it into an embedding. These embeddings can then be used for various analysis tasks, such as clustering similar stars, classifying different types of variable stars, or detecting anomalies. The contrastive curves method brings several advantages to the field of astronomy; it can efficiently handle the massive datasets generated by modern astronomical surveys, it is designed to be resilient to noise and irregular sampling (given the appropriate design of augmentations), and the embeddings produced can be used for a wide range of downstream tasks.

4.2.1 Data sample, preparation, and training

The light curves were taken from VVV light curves in the PRIMVS catalogue (Miller et al in prep). PRIMVS contains ≈ 5 million periodic variable stars with low false alarm probability.

Due to the unique nature of the VVV survey, cross matching with pre-existing periodic variable catalogues was limited. Each light curve's period was calculated via phase folding and Fourier based techniques with a false alarm probability assigned from a machine learning technique (see chapter 2). This allows for multiple periods to be calculated via fundamentally different methods and the period with the lowest false alarm probability to be used. This is performed in an attempt to remove any selection bias for periodic variable stars with previously unknown phase folded structures such as the difference between AGB and EB, even though these are known.

For each input light curve a phase was calculated from the best period. A time resolution of 0.25 days was used such that any datapoints within this distance to each other were combined and their photometric error reduced as a function of $1/\sqrt{N}$. This is required to minimise the longest light curve length within a batch; if we have a single very long light curve within a batch, the neural network will pad every curve within the batch to the length of the longest light curve. In extreme cases this requires more VRAM than is available on the GPU machine and halts training. After this reprocessing, the number of datapoints per light curve varies between 40 and 2000, with a median of ~ 150. Further cuts were used to ensure each light curve only contained reliable data-points that exclusively feature the photometry of the target star. This involved selecting for both photometric and astrometric error. A selection of the following was used:

- $m_{error} < 0.5$
- $m_{error} < 3 \times \bar{m_{error}}$
- 'ast_res_chisq' < 100
- 'chi' < 10
- 'ambi-match' $= 0$

Where 'ast_res_chisq', 'chi', and 'ambi-match' are astrometric values taken from the `DoPHOT` point spread function (PSF) fitting code. 'ast_res_chisq' and 'chi' characterise the goodness of fit for the PSF. 'ambi-match' is a boolean flag which signifies if the source appears blended with a neighbour.

We normalise our magnitudes as

$$\bar{\mathbf{x}}_m = 2\left(\frac{\mathbf{x}_m - A}{B - A}\right) - 1 \tag{4.2}$$

where apparent magnitude is denoted x_m. A is set as the 90% completeness limit of the VVV survey (16.8 mag), and B is set as the VVV survey's saturation point (12 mag). This scaling ensures that all magnitudes are roughly scaled between -1 and 1. Since the light curves are already phase folded, we can pair our magnitudes with their phases. All the phases are originally scaled between 0 and 1. Since the phase is cyclical, we embed it as a two channel vector

$$\bar{\phi} = (\sin(\tau\phi), \cos(\tau\phi)). \tag{4.3}$$

The final light curves as seen by the model have three channels: magnitude and the two channels of encoded phase. We select the following augmentations for our model:

- We want the learnt features to be invariant to the telescope's sampling schedules. To this end we apply a random datapoint deletion of our incoming sequences as an augmentor. In practice we apply dropout (Srivastava et al., 2014) at a 10% rate on our sequences.

- We also do not want the representations to be dependent on the light curve length, and so we also always apply a random crop along the time axis.
- To enforce phase invariance in the light curves we apply a randomised phaseshift on the phase folded light curves. In practice we sample a phase from $\alpha \sim \mathscr{U}(0,\tau)$, and rotate the phase channels via the trigonometric identities $\sin(\tau\phi+\alpha) = \sin(\tau\phi)\cos\alpha + \cos(\tau\phi)\sin\alpha$, and $\cos(\tau\phi+\alpha) = \cos(\tau\phi)\cos\alpha - \sin(\tau\phi)\sin\alpha$.
- The data is affected by instrumental noise. As we do not want the model to use this information in its representations, we apply a random noise addition in our augmentation pipeline. This noise is sampled from $\mathscr{N}(0,\lambda\Delta m)$, where Δm is the median magnitude error of the time series. λ is a hyperparameter. We set λ to 1 to take into account error sources that are not represented in Δm.
- We apply an amplitude jitter to the magnitude channel. This is of the form of a random resample within a flat distribution between 1-1.05 of the amplitude. Without prior classification of the light curves we cant know the expected amplitude range for the source. Instead, we use a conservative amplitude jitter as being a reasonable alternative.

We always apply random phase shift and random cropping. All other augmentations are applied at a 50% rate.

The final model is trained for 50000 iteration steps on a single NVIDIA Tesla V100. Training completes in a wall time of roughly 18.5 hours.

4.2.2 Tuning

Due to the novelty of this method it is not trivial to decide on input parameter values for the architecture and training of this network. Table 4.1 shows the hyperparameters used and their justification. As it can take on the order of days to fully train the model, two separate grid searches[1] were performed to determine all of the hyperparameters.

The grid search method used for this network was not as simple as deciding the iteration which produced the lowest loss and highest accuracy. This is because self supervised learning does not provide a loss or accuracy measure which can be used to directly determine the effectiveness of the neural network. The NT-Xent loss is used to construct a semantically meaningful embedding space, to determine the effectiveness of the network we must visually inspect the embedding space. Our metric for determining the '*best*' latent space representation was to inspect the structure of both the UMAP (McInnes et al., 2018) and PCA (F.R.S., 1901) projections. We also inspect the relationship between these projections and features from the PRIMVS catalogue.

[1] Training the network multiple times with an array of different hyperparameters

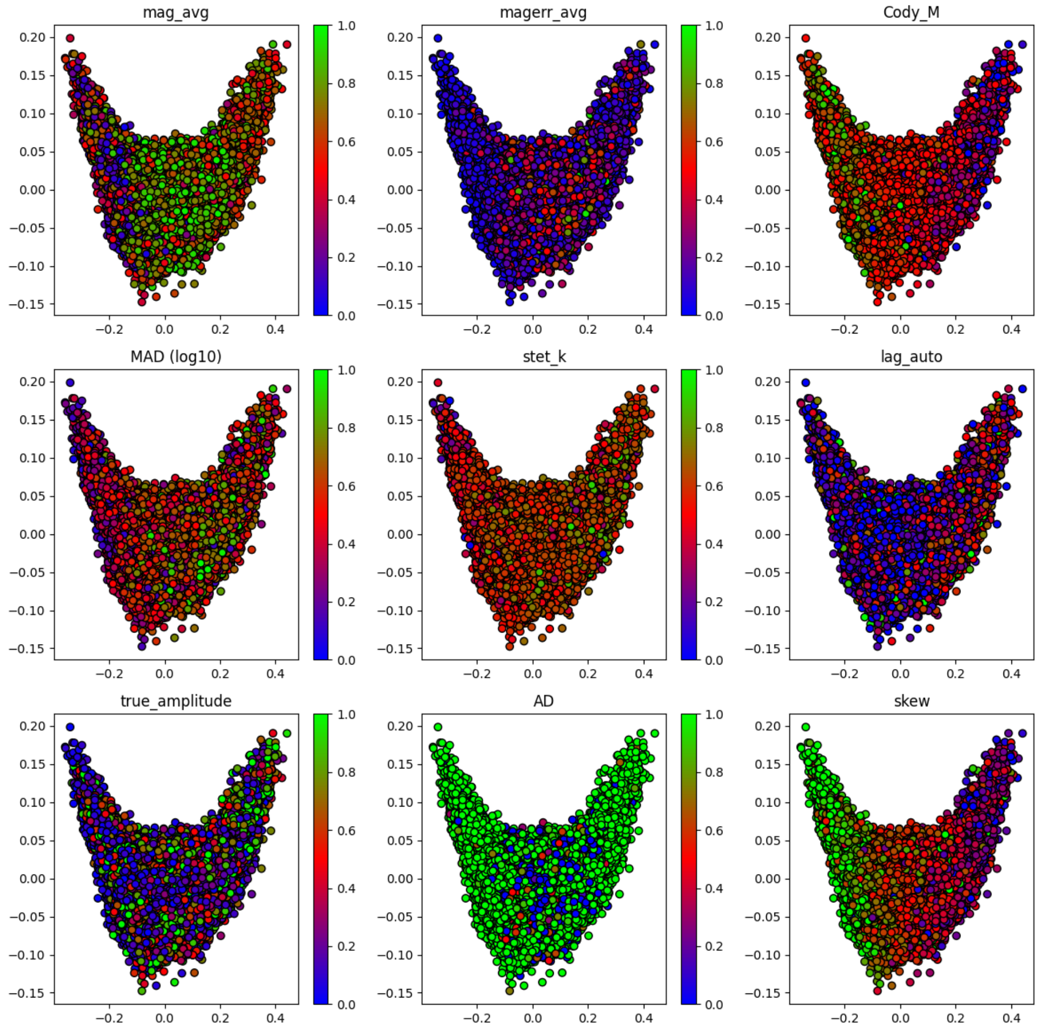


FIGURE 4.3: A PCA representation of the latent space trained with; learning rate = 0.001, tau = 0.05 and gamma = 0.7.

Figures 4.3 -4.4 show the variety of projections we receive with relatively minimal changes to the input parameters. Each subplot is colour coded with the normalised values from the PRIMVS catalogue, these are (from top left to bottom right): Average magnitude, Average magnitude error, M (Cody et al., 2014), Median Absolute Deviation, Stetson-k index (Stetson, 1996), Lag-1 autocorrelation, Amplitude, Anderson Darling (Anderson and Darling, 1952) and Skew. Each of these are described in section3.6 in chapter 3.

The observed sensitivity in latent space with respect to the selected hyperparameters means it is likely we have not selected the most optimal values. However, the values we have selected produce a semantically meaningful latent space projection and further tuning will require a more intelligent approach.

The first grid search ”Grid search # 1” was used to determine parameters that do not specifically pertain to the training loop - drop out rate and the number of hidden dimensions. The drop out

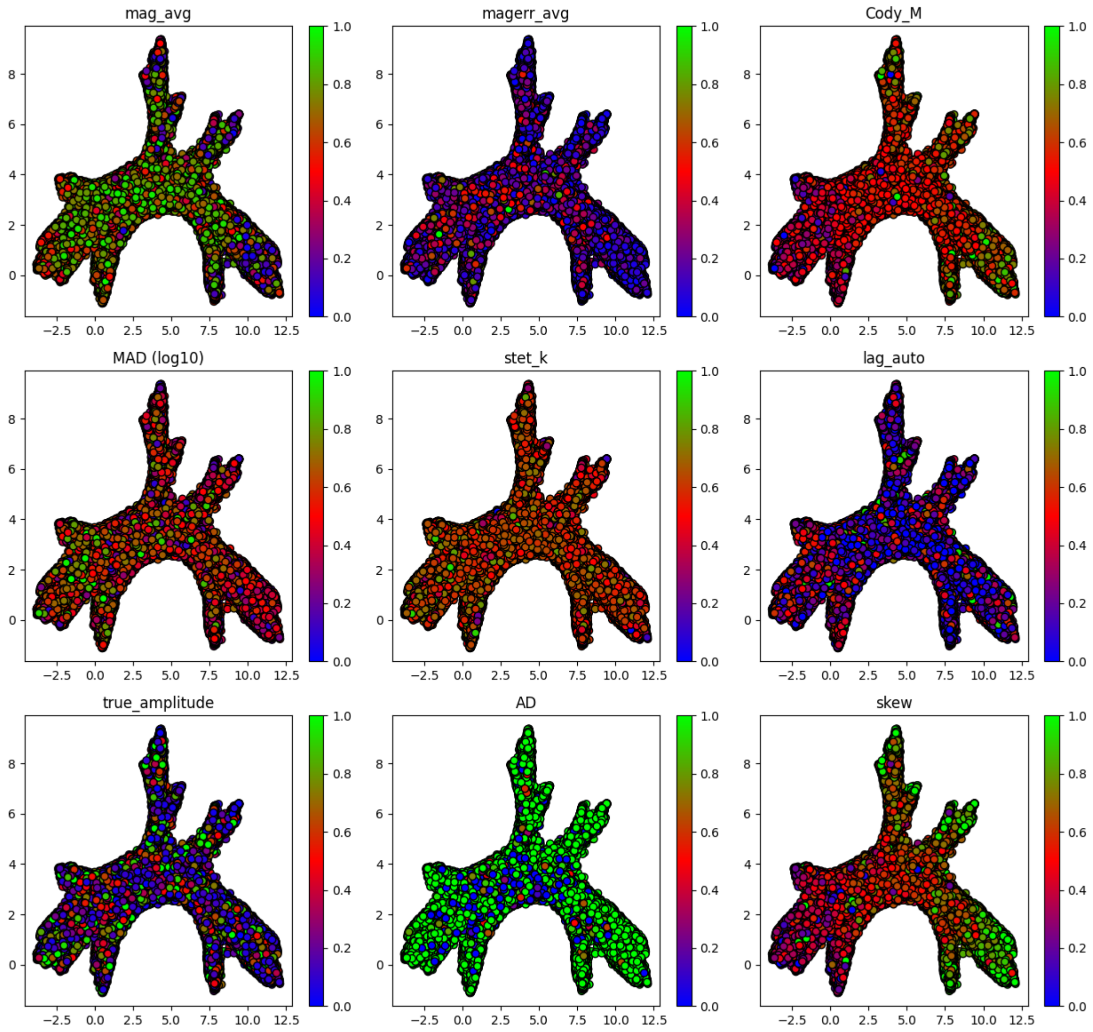


FIGURE 4.4: A PCA representation of the latent space trained with; learning rate = 0.01, tau = 0.05 and gamma = 0.9

TABLE 4.1: Tuning parameters and their justifications

Hyperparameter	Value	Justification
Batch Size	4096	GPU memory limited
Drop out rate	10%	Grid search #1
Hidden Dimensions	64	Grid search #1
Learning rate	0.0001	Grid search #2
Tau	0.05	Grid search #2
Gamma	0.7	Grid search #2
Output Dimensions	64	Final tuning

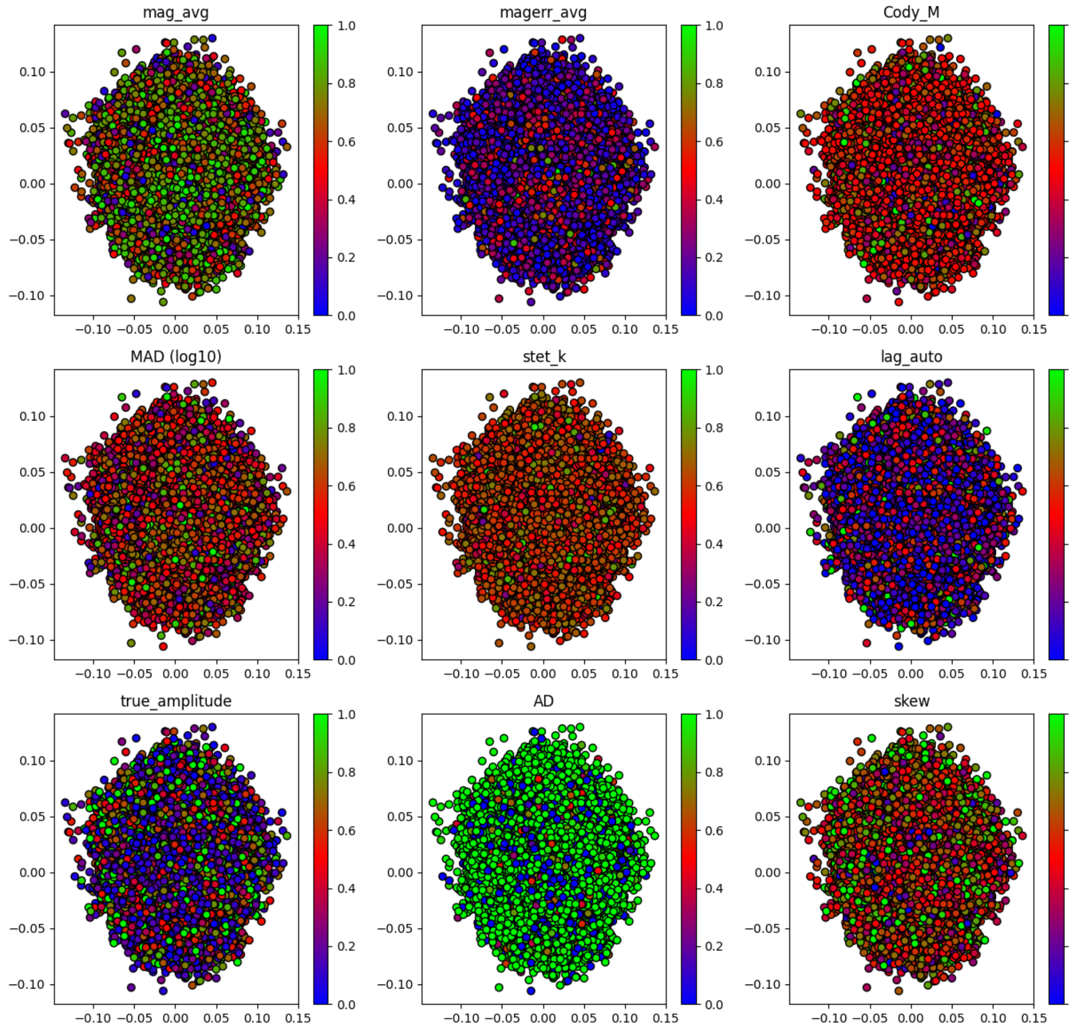


FIGURE 4.5: A PCA representation of the latent space trained with; learning rate = 0.001, tau = 0.01 and gamma = 0.7

rate determines the probability of any point in a light curve being removed. This has proven to be an effective technique for destroying highly correlated relationships between neurons(Hinton et al., 2012). The hidden dimensions (amount of neural layers between input and output) were chosen as the smallest power of 2 which did not visually impact latent space clustering.

During the first grid search values of Learning rate $= 0.01$, Tau $= 0.5$ and Gamma $= 0.9$ were used. The second grid search 'Grid Search #2' was used to determine the hyperparameters used for the training loop. Where 'Tau' is the 'temperature parameter' from the NT-Xent loss, it is analogous to the learning rate with a larger value amplifying gradients through the network. 'Gamma' determines the rate at which the learning rate decays during training. The number of output dimensions was determined as the final parameter. This value was iteratively halved from 256 until the PCA and UMAP representations of both hidden states (h) and latent representations (z) noticeably declined in complexity. Interestingly, this appeared to be 64 dimensions, the same as the hidden dimensions.

4.3 Results

For the classification of stellar light curves, the utilisation of both hidden states (h) and latent representations (z) proves to be advantageous. Hidden states encapsulate the temporal dynamics inherent to light curves. This is crucial for capturing patterns such as periodicity and trends over time. Latent representations are hidden states passed through a feed-forward network, in our case a projection layer consisting of linear transformations, `LeakyReLU` activation, and dropout regularisation. Latent representations offer a condensed version of the input data, with the goal of emphasising the key features that are essential for classification.

Given the distinct characteristics of stellar light curves, the combined use of h and z can significantly improve model efficacy. h leverages the sequential nature of the data to capture dynamic changes, while z distills this information into a feature-rich representation ideal for classification.

Figures 4.6,4.7,4.8, and 4.9 show the PCA representation of the hidden states with the colour axis representing different features from the PRIMVS catalogue.

Figure 4.6 shows the PCA representation of the hidden states as a function of VVV colours. There is a slight indication of a correlation with clumps of colours loosely forming. This is a weak correlation however and the correlation is dominated by noise. The plot demonstrates the model's potential to identify previously unlabelled stellar classes that feature a correlation with VVV colours, despite this colour information not being a factor in the latent space construction. This ability to correlate with known physical properties, despite the model not being explicitly trained on them, hints at a correlation with stellar class. This suggests that the clustering based on light curve morphology is likely real because morphology is related to class and, independently, colour is also related to class.

Figure 4.7 shows the PCA representation of the hidden states as a function of basic statistics for the light curves: skewness, kurtosis (in $\log_{10}$), amplitude, and period. We observe a strong correlation with skewness, indicating that the contrastive curves method is effectively creating representations based on the shape of the light curve. This is a crucial validation that our model is sensitive to morphological features, which are essential for distinguishing different types of variable stars. In contrast, we do not see a strong correlation with kurtosis, which may be due to the challenging nature of representing the distribution shape accurately. Kurtosis measures the tails of the distribution, and its weaker correlation might indicate that the model does not prioritise these features as strongly as skewness or other statistics. Encouragingly, there is a visible correlation with both amplitude and period, which are commonly used together to form a Bailey diagram to aid in stellar classification. This suggests that the model captures key periodic characteristics and the extent of brightness variation in the light curves. The independence of these correlations from skewness further supports the robustness of the model, demonstrating

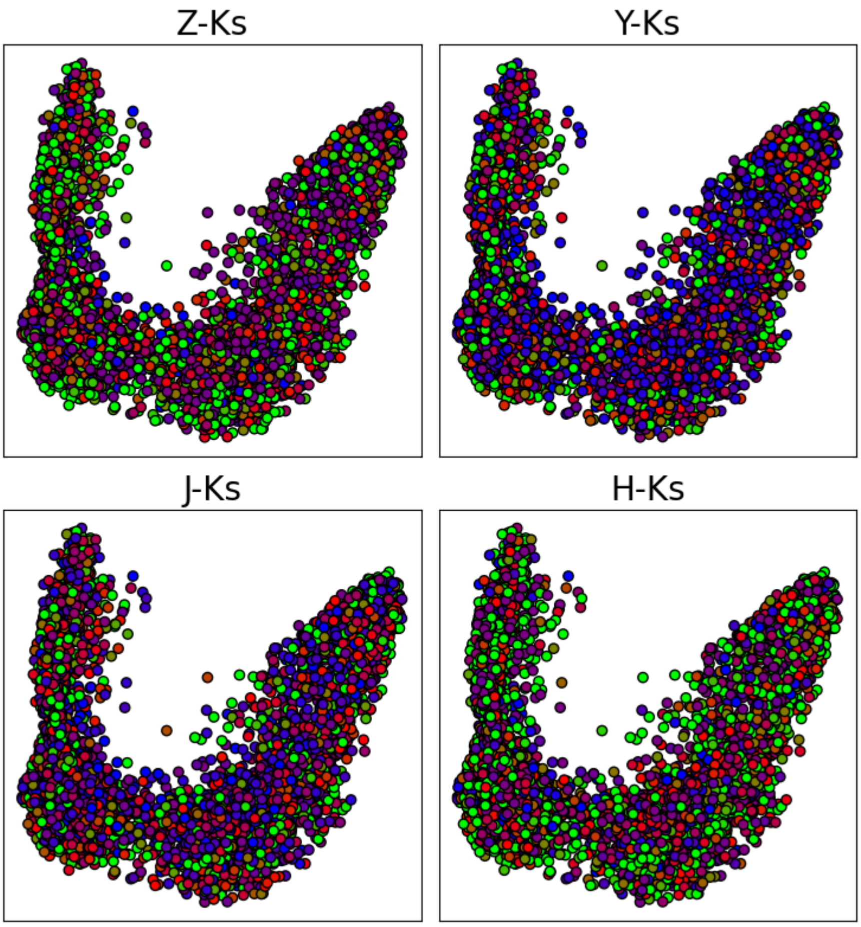


FIGURE 4.6: A 2 dimensional PCA representation of the hidden states (h) colour coded with respect to VVV colours. The plot demonstrates the ability of the model to identify previously unlabelled stellar classes which feature a correlation with VVV colours, despite this colour info not being a factor of the latent space construction.

its ability to consider multiple dimensions of variability simultaneously. This independence hints at the deeper complexity of the latent space representations, indicating that the model can discern and encode various aspects of the light curves. By clustering light curves based on shape, amplitude, and period without explicit input on these statistics, the model demonstrates its capability to uncover intrinsic patterns in the data. This alignment with known statistical properties underscores the reliability of the model's representations, providing confidence in its application to other astronomical datasets.

Figure 4.8 shows the PCA representations as a function of more statistical features extracted from the light curve (see chapter 3 section 3.6). We see a relatively strong correlation with every feature shown. A.M Cody's 'M' value measures the asymmetry in the light curve, with higher values indicating more pronounced asymmetry, a key parameter for identifying eclipsing binaries. The median buffer range percentage captures the proportion of points within a small range around the median magnitude. The range of cumulative sum assesses the overall variability, with larger values indicating greater variability. The maximum slope identifies the steepest

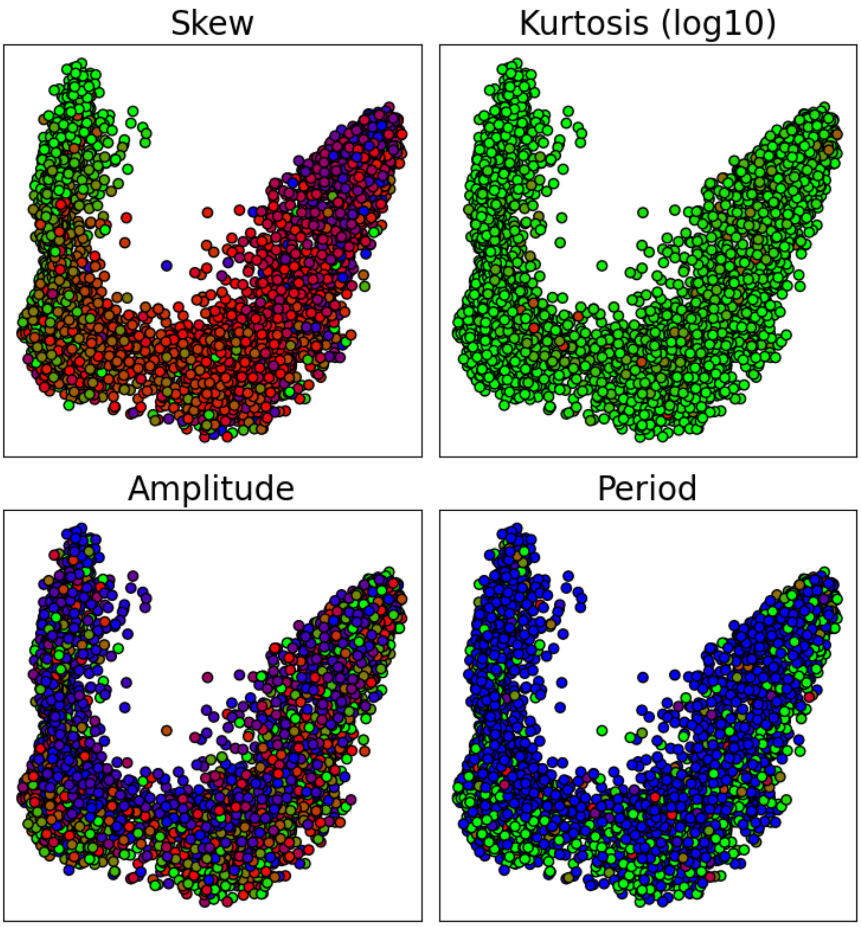


FIGURE 4.7: A 2 dimensional PCA representation of the hidden states (h) colour coded with respect to Skew, Kurtosis, Amplitude and Period (from top left to bottom right).

change in brightness, highlighting rapid variability. The independent correlations observed in the PCA representations further highlight the model's ability to capture more nuanced details of light curve variability.

Figure 4.9 shows the PCA representations as a function of features that might indicate data quality or non-desirable correlations. Ideally, the latent space should exhibit minimal dependence on factors such as Galactic position, apparent magnitude, or photometric uncertainty. In these representations, we observe no clear correlation with Galactic position, indicating that the model is effectively capturing intrinsic properties of the light curves rather than spatial biases. However, a correlation with both magnitude and its associated error is evident. The augmentations used in this method aim to mitigate these dependencies, but they are fundamentally constrained by the assumption that photometric uncertainty is uncorrelated scatter. This assumption is likely not always true, leading to residual correlations. Addressing this limitation, future iterations of this work should focus on developing more sophisticated models for photometric uncertainty. Such models would account for correlated noise and other systematic effects, enhancing the robustness of the latent space representations by removing unwanted dependencies.

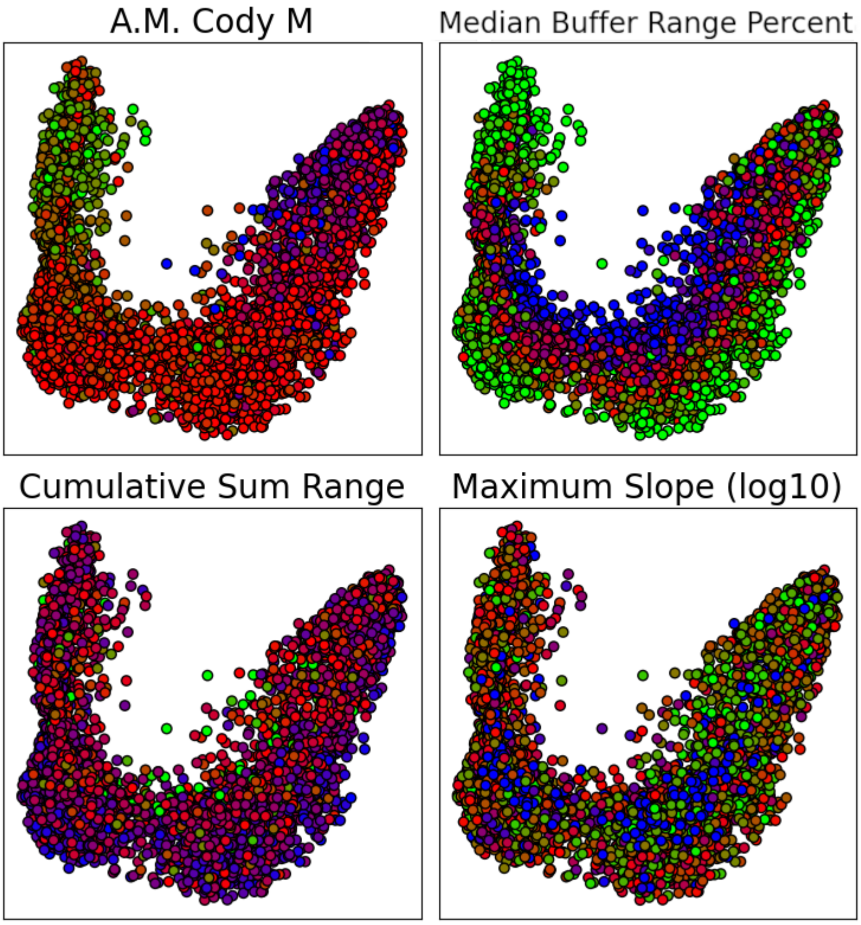


FIGURE 4.8: A 2 dimensional PCA representation of the hidden states (h) colour coded with respect to 'M', the median buffer range percentage, the range of cumulative sum, and the maximum slope (from top left to bottom right).

Figure 4.10 shows the PCA representation of the latent representations, z. It can be seen that these representations provide less useful representations of the latent space. This is likely because the latent representations are the hidden states which have been passed through an auto-encoder. Ablative testing has shown 64 dimensions to be the minimum at which the auto-encoder still preserves all apparent semantically useful information. It makes sense that further reducing this information via PCA is not useful. This does not mean the latent representations are useless; rather, they are hard to properly visualise in lower dimensions.

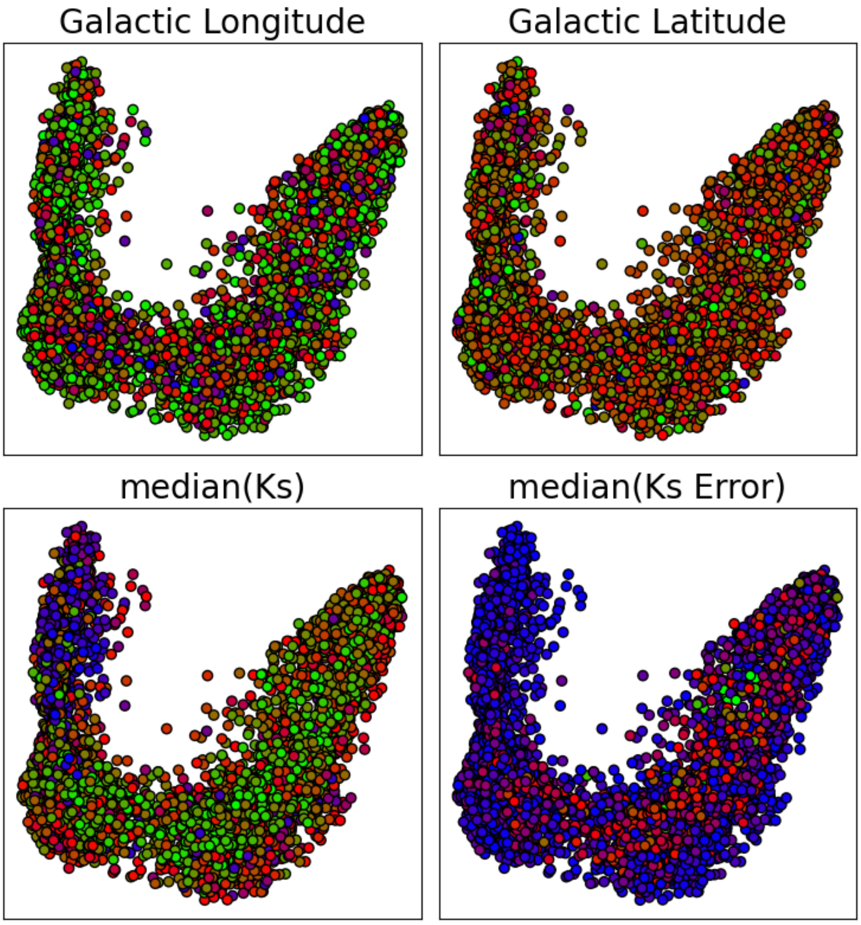


FIGURE 4.9: A 2 dimensional PCA representation of the hidden states (h) colour coded with respect to Galactic latitude, longitude, madian magnitude and median magnitude error (from top left to bottom right).

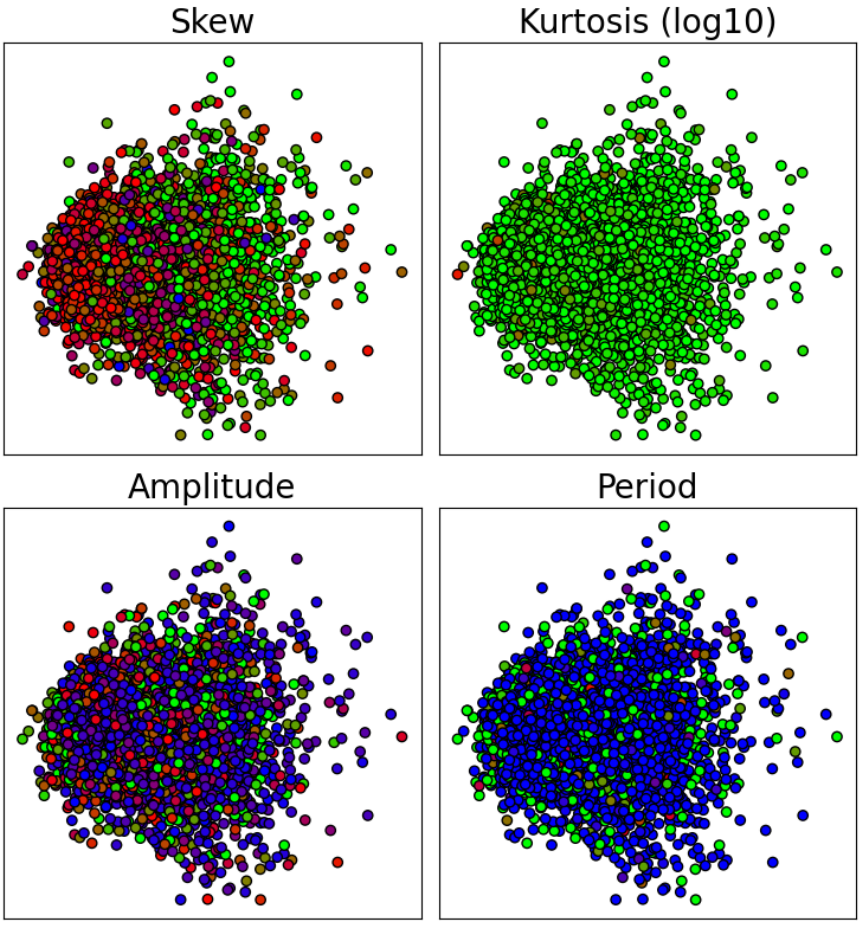


FIGURE 4.10: A 2 dimensional PCA representation of the latent representation (z) colour coded with respect to Skew, Kurtosis, Amplitude and Period (from top left to bottom right).

4.3.1 Latent Space Exploration

We can use the classifications obtained in chapter 3, section 3.7.3 as a way of verifying this method. The method in chapter 3 is a decision tree based on a Gaia training set and extracted astrometric, colour and time series features. This means that it is almost entirely independent of the contrastive curves method. Figure 4.11 shows two views of the 3 dimensional PCA representation of the hidden states. A clear separation can be seen between stellar class predicted by the decision tree in the latent space. This further suggests that we are indeed generating semantically meaningful representations of phase folded light curves.

Given that we see distinct clusters of labelled classes in figure 4.11, we can look for unlabelled objects within these clusters. We are effectively using these externally labelled classes to trace the classes from our contrastive curves representations. We expect the unlabelled objects within each cluster to be of the class that the cluster represents. The centre of each cluster is defined as the mean value (in terms of the 3-dimensional latent space projection) of all sources with a probability > 0.7, entropy < 0.2, and confidence metric > 0.9 (see chapter 3, section 3.7.3). Figures 4.12, 4.13, 4.14 and, 4.15 show the phase folded light curves of the 4 nearest neighbours to the centre of the Eclipsing Binary, RR Lyrae, Cepheid, and Ellipsoidal clusters, respectively. These figures highlight the ability of the model to accurately classify and identify new members of these stellar classes.

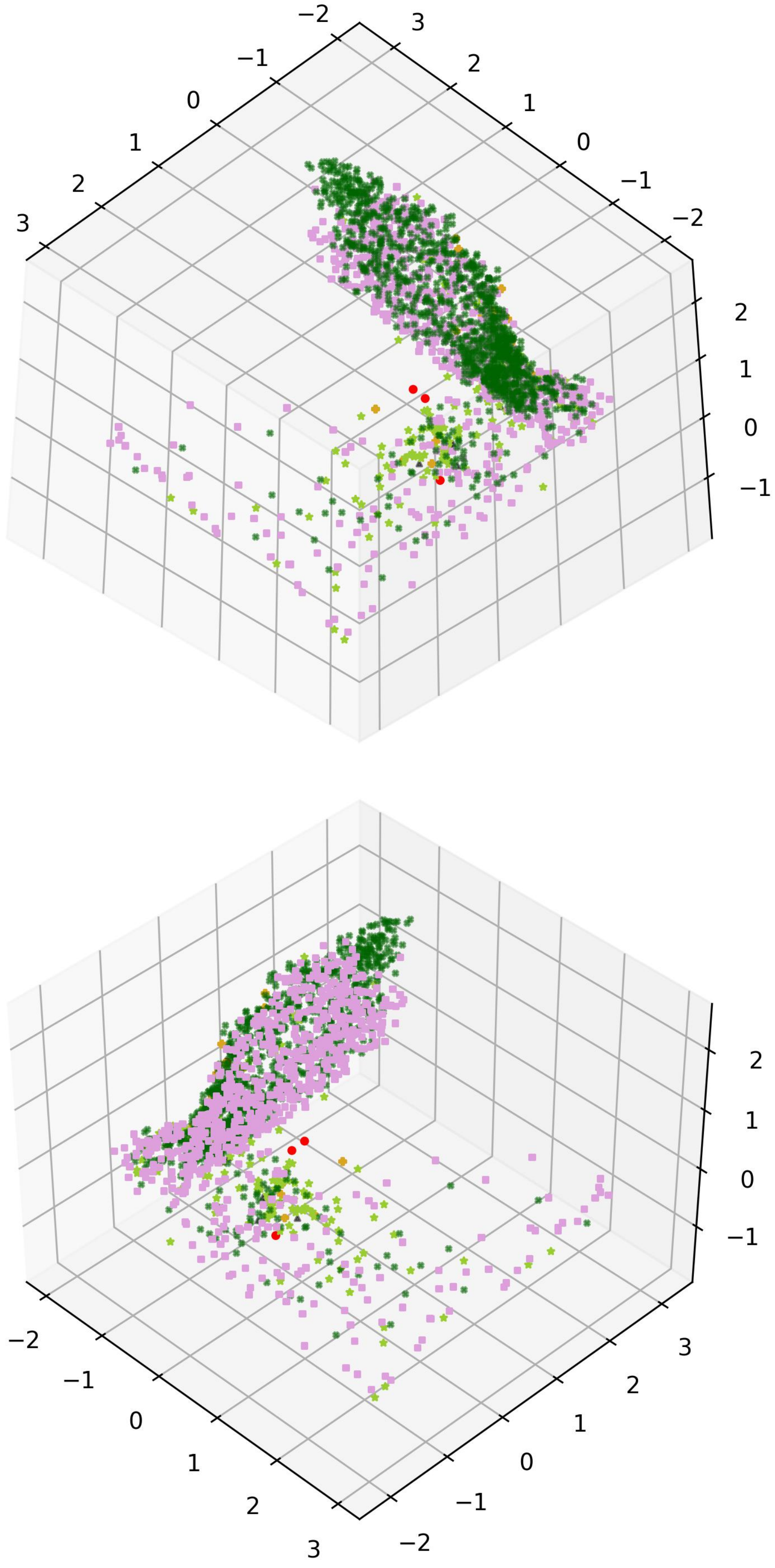


FIGURE 4.11: Two different points of view for the 3 dimensional PCA representation of the hidden states (h). There colours here are the same as figure 3.22. Where dark green is eclipsing binary, pink is long-period variables, yellow is rr lyrae, red is cepheids, and light green is ellipsoidals.

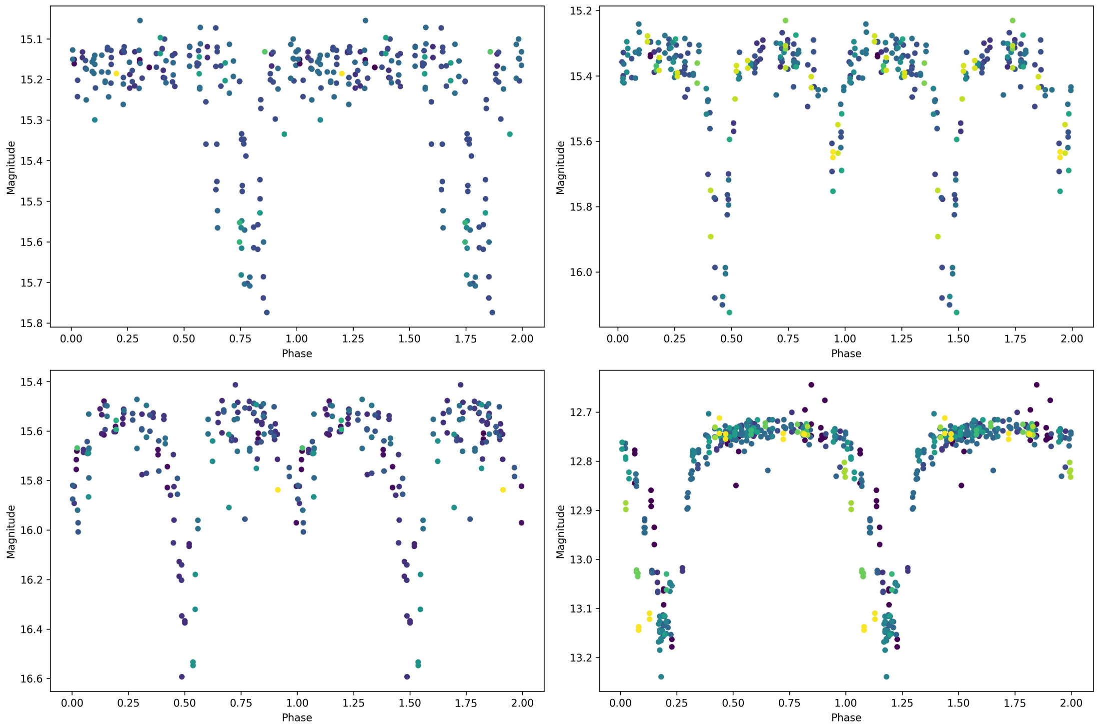


FIGURE 4.12: The phase folded light curves of the 4 nearest neighbours to the centre of the Eclipsing Binary cluster.

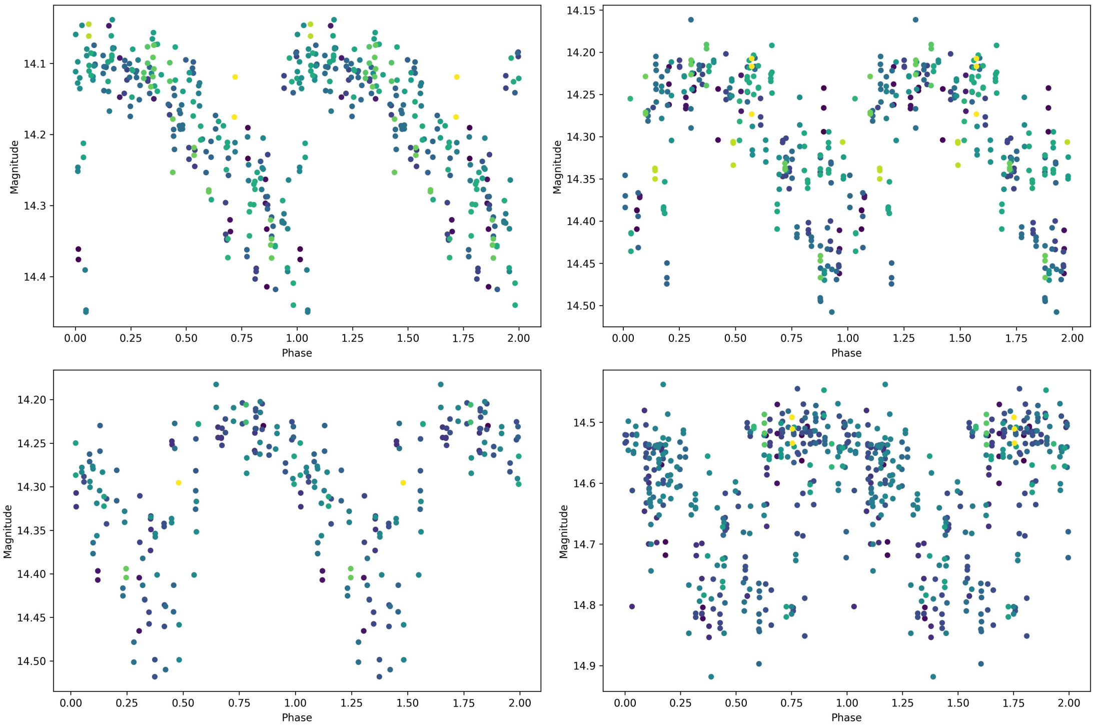


FIGURE 4.13: The phase folded light curves of the 4 nearest neighbours to the centre of the RR Lyrae cluster.

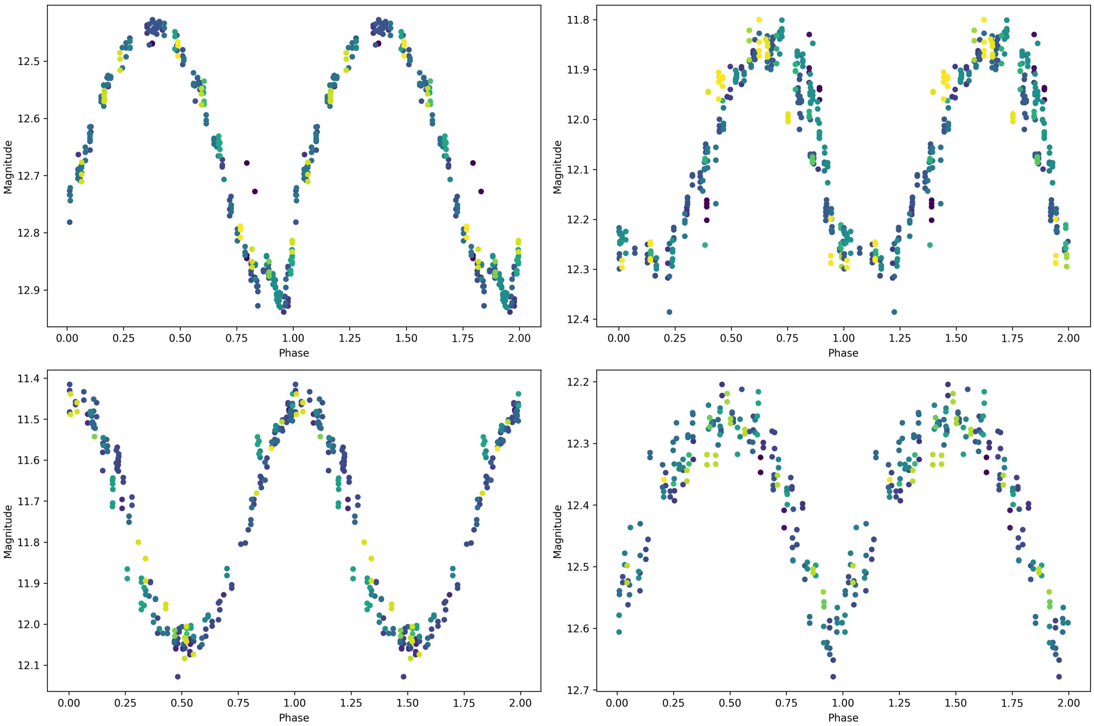


FIGURE 4.14: The phase folded light curves of the 4 nearest neighbours to the centre of the Cepheid cluster.

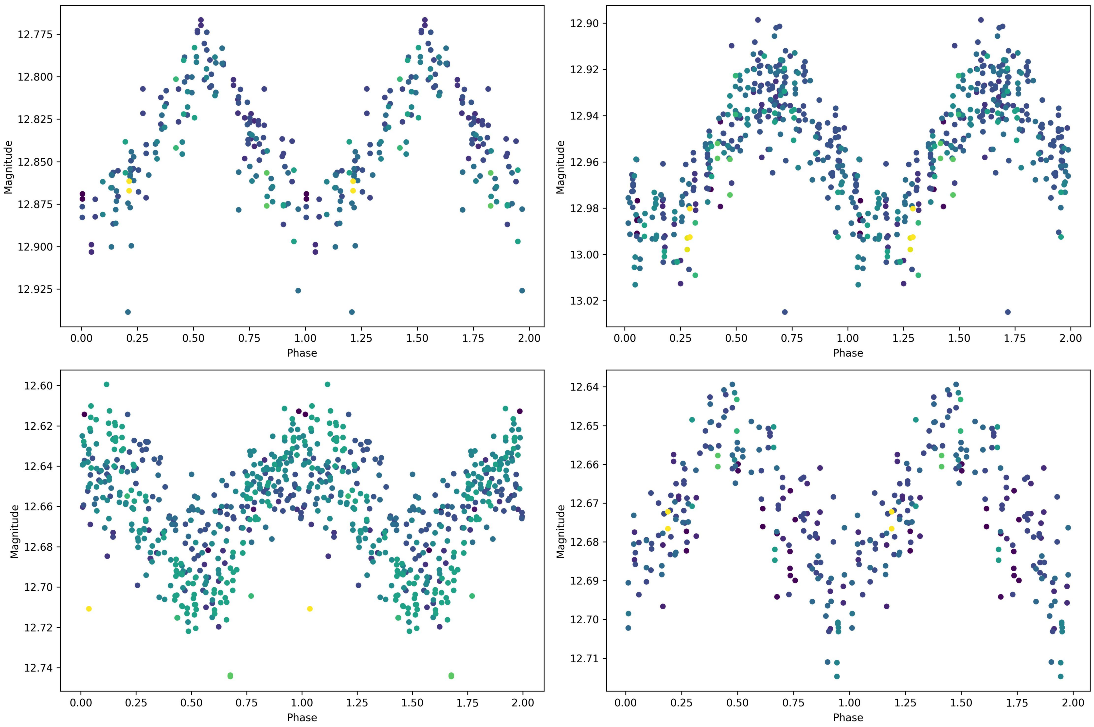


FIGURE 4.15: The phase folded light curves of the 4 nearest neighbours to the centre of the Ellipsoidal cluster.

4.4 Conclusions

The paper demonstrates the effective use of a novel neural network architecture, which leverages contrastive learning with a gated recurrent neural network backbone, to generate semantically meaningful representations of stochastically sampled time series data. This approach proves to be effective for analysing astronomical time series data, capturing the complex, dynamic behaviours characteristic of variable stars.

Training the model presented unique challenges, primarily due to the novelty of the approach and the complexity of the time series data. Through extensive experimentation, including two rounds of grid searches, some optimal hyperparameters were identified.

The analysis presented here is by no means exhaustive and only seeks to prove the efficacy of this method. A future work into this would be to combine these representations with orthogonal features (such as colour and astrometry) via an autoencoder, a continuation of the work seen in chapter 3 section 3.7.2.

We can also use these features in our earlier classifications, seen in chapter 3 section 3.7.3. We have shown with figures 4.11- 4.15 that there is good agreement between the Gaia trained decision tree based classification and the contrastive curves method. It follows that using the output of contrastive curves in the decision tree would likely improve classification accuracy.

This work showcases the potential of contrastive learning models to revolutionise our understanding of astronomical time series data. By effectively capturing the essence of variable stars in fixed-length embeddings, this approach opens up new possibilities for automated classification and analysis in the era of Big Data in astronomy.

Chapter 5

Future Work

There are multiple projects that I did not finish before the end of my PhD. I highlight the two most exciting here as well as my research proposal, which was used to highlight work I intend to undertake during my post doctoral position.

5.1 VVV-Gaia Distances

Gaia is a space-based observatory launched by the European Space Agency (ESA) with the primary mission of compiling a comprehensive astrometric catalogue of over a billion sources. The satellite operates by continuously scanning the sky with its two optical telescopes. Due to the location, observing pattern, and design of Gaia, it excels in producing notably precise astrometric measurements. Gaia's observational campaign is distinguished by its high resolution and depth, with the goal of achieving parallax measurements with microarcsecond accuracy for stars as bright as $G \approx 3$ mag down to $G \approx 21$ mag. Gaia DR3 provides astrometric solutions for around 1.46 billion sources(Gaia Collaboration et al., 2023). The median parallax uncertainties are $0.02 - 0.03$ mas for $G < 15$ and $0.5 - 1.3$ mas at $G > 20$ mag. Despite its robust capabilities in optical astrometry, Gaia encounters limitations when observing regions suffused with interstellar dust. The median parallax uncertainties are $0.02 - 0.03$ mas for $G - Ks < 2.5$ and > 1 mas at $G - Ks > 7.5$ mag. Figure 5.1 shows the 2D histogram of 10,000 non-variable lightcurves in the VVV survey, cross-matched to Gaia. The strong correlation between colour and parallax error can be seen.

Star-forming areas, known for their dense dust concentrations, pose a significant challenge to optical observations. In these environments, the optical signals that Gaia relies on are absorbed and scattered by dust particles, leading to a decrease in the effectiveness of its measurements. This reduction in observational efficacy is particularly critical when looking at YSOs and differentiating their contaminating AGB counterpart. As AGB stars exist in a very similar colour

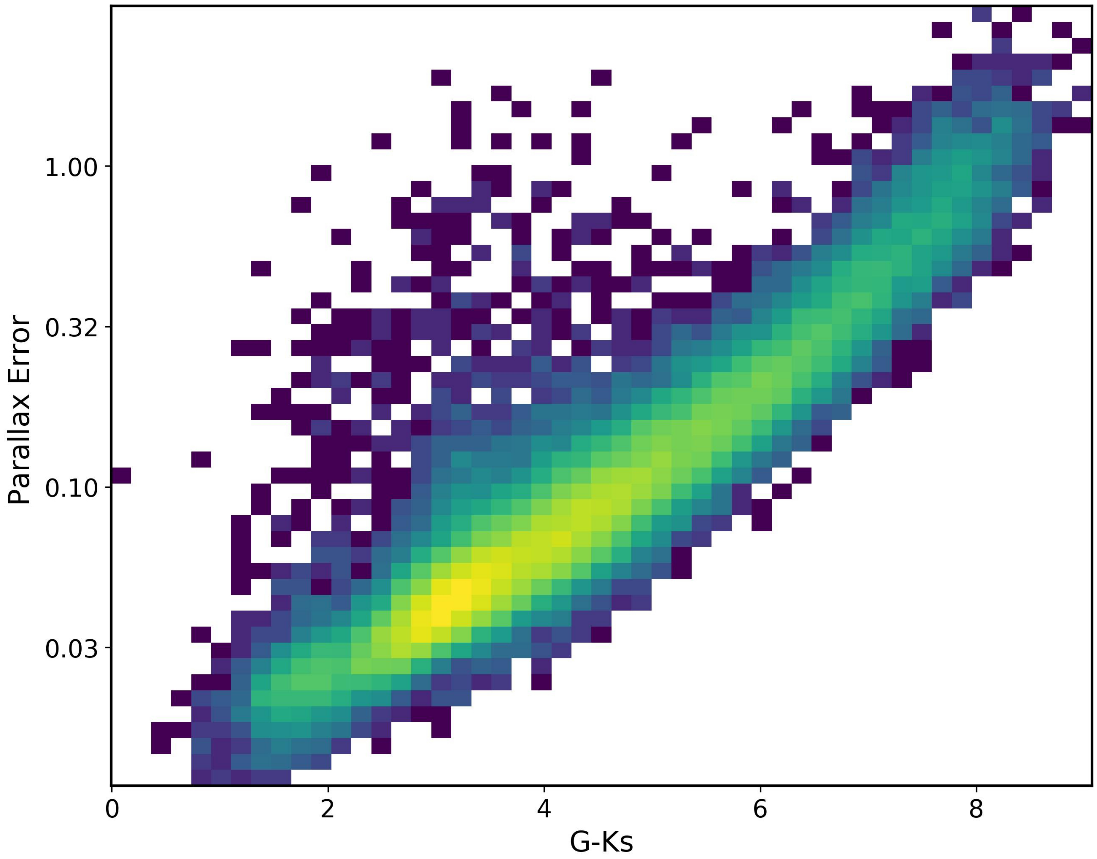


FIGURE 5.1: A 2D histogram of Parallax Error verses $G - Ks$. The strong relationship between colour and parallax error is evident.

space (dust driven IR excess) it is essential we use other features to resolve degenerecies between the two. Due to the nature of star forming regions, YSOs found in these locations are of similar distances to each other. However, AGB stars are not constrained to star forming regions objects in the background that can contaminate star formation studies. It follows that parallax measurements are critical for this disentanglement of these sources. VVV is a survey conducted using the VISTA telescope that focuses on the southern viewable Galactic disk and bulge in the near-infrared spectrum. VIRAC is the astrometric solution for VVV, calibrated to align with Gaia's measurements. VVV's near-infrared capabilities allow it to observe areas that are heavily obscured by dust. The similarities between Gaia and VVV's resolution makes it ideal for improving astrometry. A Multilayer Perceptron (MLP) (see figure 1.6) is used to combine VVV and Gaia data. The MLP's role is to facilitate the merging of astrometric data from both surveys. This allows us to increase the quality of parallax measurements where Gaia's optical observations may be obstructed by dust but VVV's are not.

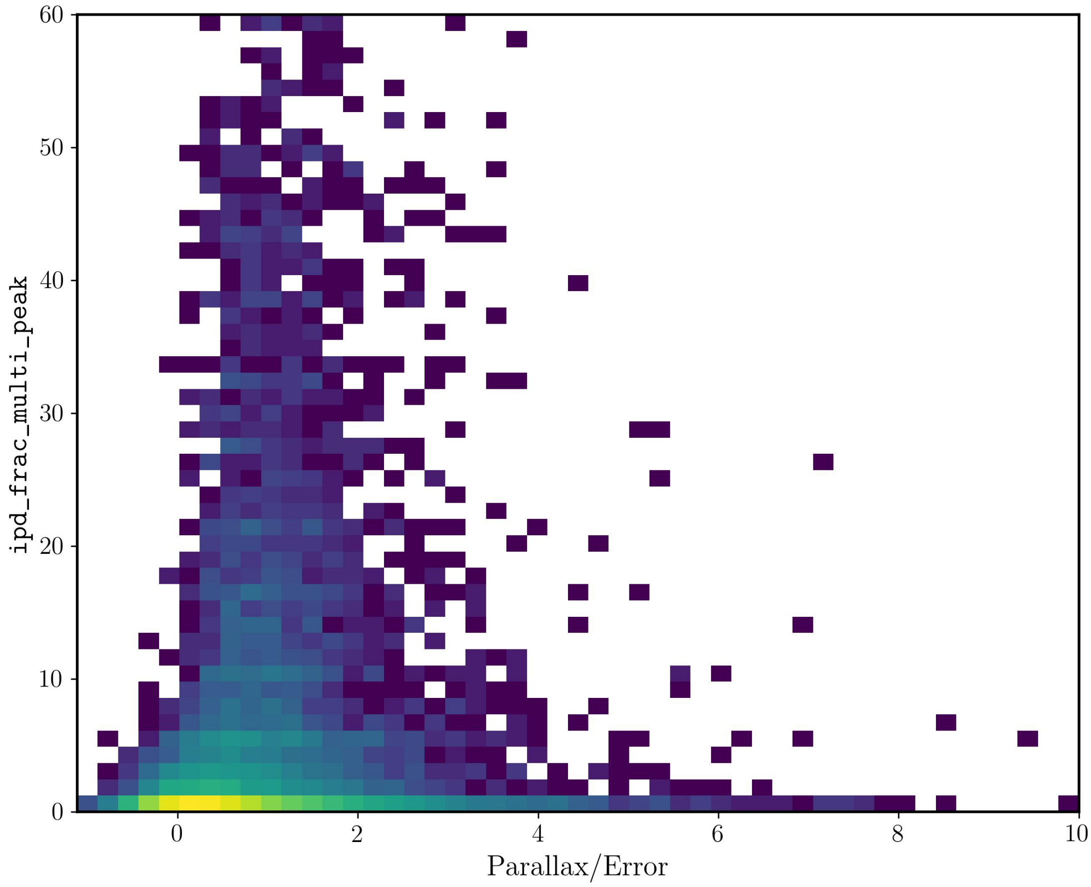


FIGURE 5.2: A 2D histogram of ‘Parallax/Error’ versus ‘`ipd_frac_multi_peak`’. Where ‘`ipd_frac_multi_peak`’ is the fraction of source extractions which feature multiple peaks in their psf. The majority of sources have a multi peak fraction $< 10\%$ and the majority have a Parallax/Error< 2

5.1.1 Training

We use a training set of non-variable VVV sources, as these sources will have more reliable astrometry than variable sources (Stassun and Torres, 2021). Non-variable sources are identified after the light curve is cleaned in the same way as seen in chapter 3 section 3.4. A selection of the following is used to define non-variable VVV sources:

- $\Delta m < 0.1$
- $\Delta m/m_{error} < 1.1$

We cross-match these VVV non-variable sources to the Gaia catalogue. A grid search was used to determine the optimal selections to use for the training data.

There is a trade off between data quality and data size, it is not a trivial selection in instances where data size is significantly impacted by quality selections.

gaia pmra, pmdec	Gaia proper motion for right ascension and declination
phot bp rp excess factor corr	BP/RP excess factor
ra/dec parallax corr	Correlation between right ascension/declination and parallax
parallax pmdec/pmra corr	Correlation between proper motion in right ascension/declination and parallax
pm	Total proper motion
pmra pmdec corr	Correlation between proper motion of right ascension and declination
ra dec corr	Correlation between right ascension and declination
radial velocity	Radial velocity
bp g, bp rp, g rp	BP-G, BP-RP & G-RP colours
grvs mag	Integrated Grvs magnitude
vvv parallax	VVV/VIRAC parallax
vvv pmra, pmdec	VVV/VIRAC proper motion for right ascension and declination
z med mag-ks med mag	Median z band magnitude - median ks mag
y med mag-ks med mag	Median y band magnitude - median ks mag
j med mag-ks med mag	Median j band magnitude - median ks mag
h med mag-ks med mag	Median h band magnitude - median ks mag
l	galactic longitude
b	galactic latitude

TABLE 5.1: Table showing each of the features used in the training of the MLP. 'ra/dec parallax corr' and 'parallax pmdec/pmra corr' were shortened for formatting reasons.

Figure 5.2 shows the distribution of 'Parallax/Error' verses 'ipd_frac_multi_peak' for the training set. Where 'ipd_frac_multi_peak' is the percentage of successful image parameter determination windows with more than one peak (i.e. Gaia's version of having multiple peaks in the PSF). Ideally the majority of sources will have a Parallax/Error> 2. However, this is not the case and so the quality selection for these two parameters are determined via the aforementioned grid search. The grid search showed a sensitivity to hyper-parameters and data quality, thus it is likely we do not have the optimal parameters for this network.

The MLP used for this method has six hidden layers. These expand the dimensions to 512 and are then iteratively halved through each layer to 16 dimensions before the output of 1 dimension (the predicted parallax). Table 5.1 shows the features that were used to train the MLP.

Each of the features were selected so that combined they should form a unique set that describes their location within the Milky Way. The features used in the training process were selected based on their relevance to the astrometric properties being modelled. Features that provided unique information about the spatial distribution of stars within the Milky Way were prioritised. However, the feature selection process was primarily guided by domain knowledge rather than automated selection methods. This feature set is not conclusively the best combination of features. Future work could involve applying feature importance ranking techniques, such as permutation importance or SHAP values (Lundberg and Lee, 2017), to refine the feature set.

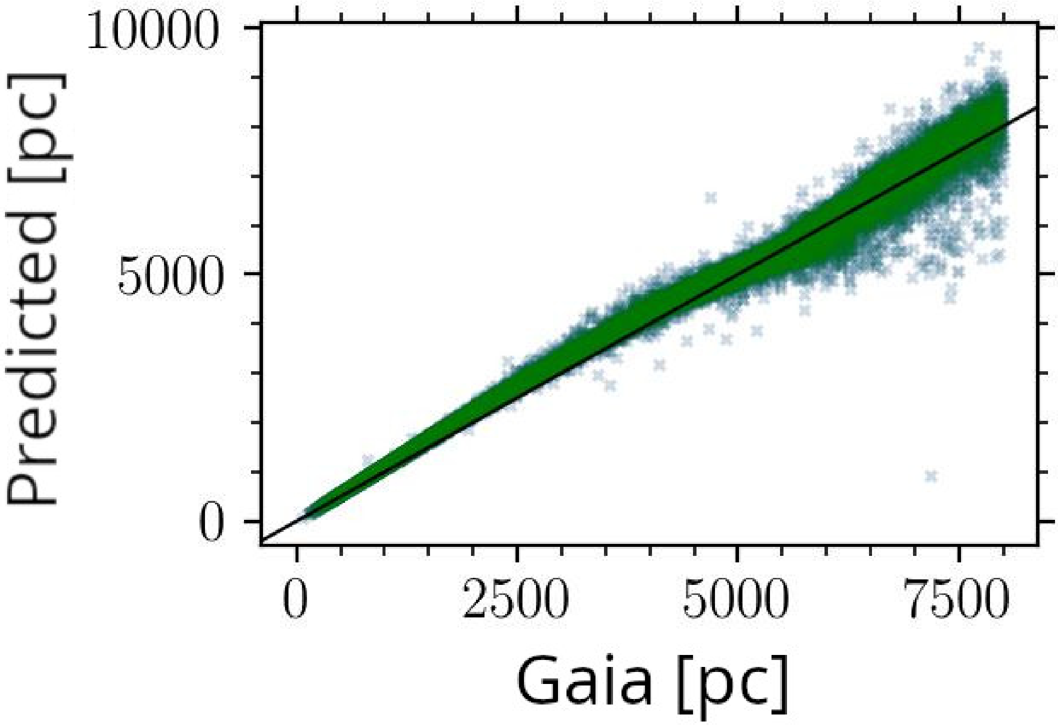


FIGURE 5.3: Predicted distance versus true distance for sources nearer than the galactic centre.

5.1.2 Initial Results

Figure 5.3 shows a scatter plot comparing predicted and true distances for sources nearer than the Galactic centre. The majority of points can be seen following the line of equality. This suggests that for stars nearer than the Galactic Center, the predictions reasonably match the true values. However, there is a scatter around this line which increases with distance. This is likely a result of increasing uncertainty in measurements for training. The limited size of the training set, especially at larger distances, highlights potential issues regarding the model's ability to generalise to unseen data. The scarcity of data points in this regime means that the model may not have learned the underlying distribution effectively, leading to poor performance when predicting distances beyond the Galactic center. To mitigate this, future work could explore data augmentation techniques or the inclusion of synthetic data to increase the training set size and diversity, thereby improving the model's generalisation capabilities. The spread of points might also be due to noise/impurities in the data or limitations in the model's capacity to capture the system.

In figure 5.4 it is apparent that this method struggles significantly with distances $> 10^5$ parsec. This is expected as there is a decrease in signal-to-noise as a function of distance as well as a decrease in sample size. Less than 2% of the training set exists at this range. The observed scatter could also be influenced by the inherent limitations of the model, such as overfitting to

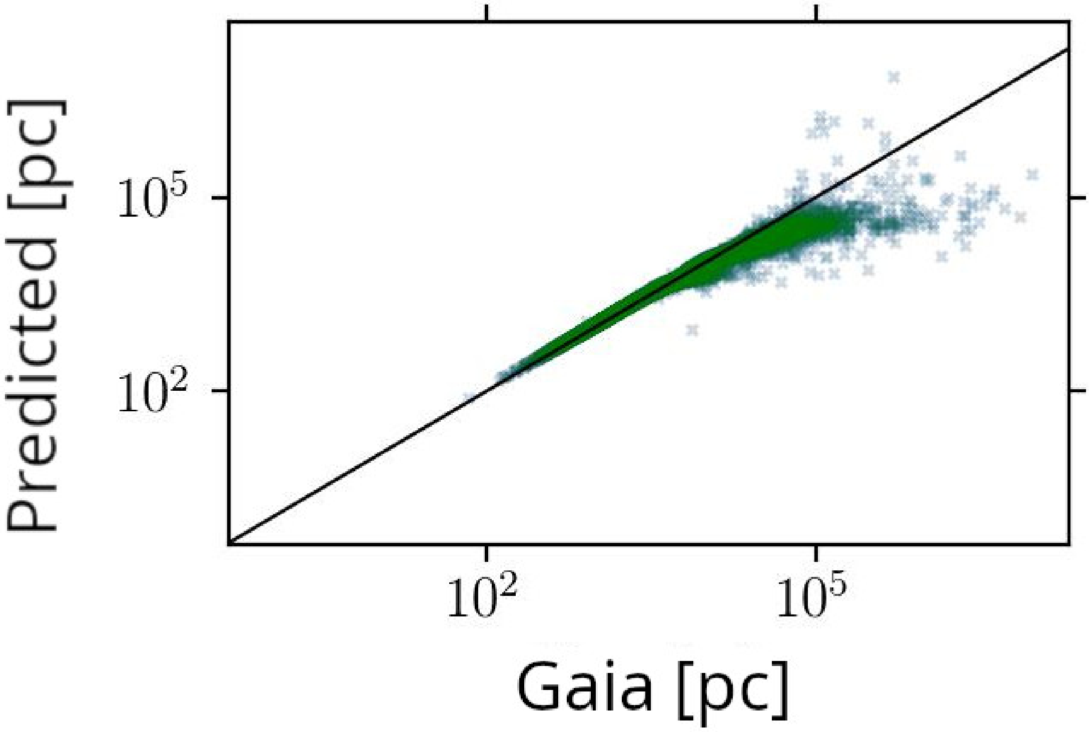


FIGURE 5.4: Predicted distance versus true distance in log scale. All sources are used here.

the training data or underfitting due to the complexity of the model relative to the data. Additionally, the input features used may not fully capture the resolve to a precise distance, leading to inaccuracies in the predictions. Future work will involve refining the model architecture and feature selection to address these issues, potentially incorporating more advanced techniques such as feature engineering or regularisation to improve model robustness.

While the initial results are promising, the limitations highlighted suggest that caution should be exercised in interpreting the model's predictions, particularly at larger distances. In this projects current state it is most useful for nearby star forming regions, most of which already have reliably distance measurements. Further refinement of the model and feature set is expected to yield more reliable results.

5.2 PRIMVS/VVV DDPM

Denoising Diffusion Probabilistic Models (DDPMs)(Ho et al., 2020) are a class of generative models that have gained significant attention in the field of machine learning and computer vision for their ability to generate high-quality, diverse samples that closely resemble the training data.

These models work by gradually converting a sample from a simple distribution, such as Gaussian noise, into a sample from the target distribution, such as astronomical light curves. This is achieved through a process that is conceptually the reverse of diffusion, which gradually adds noise to the data. The fundamental idea behind DDPMs is a forward diffusion process that gradually adds noise to the data over a series of steps until the original data distribution is transformed into a simple, known distribution. This process is reversible, and the model learns the reverse diffusion process, which iteratively denoises the data, recovering a sample from the target distribution from the noise. The reverse process is conditioned on the noisy data and is trained to minimise the difference between the original data and the reconstructed data at each step of the reverse diffusion. The iterative nature of the process allows for fine control over the generation process, including the ability to generate samples with specific attributes by manipulating the conditions under which the reverse diffusion is performed. DDPMs have demonstrated remarkable performance in generating realistic images (Sasaki et al., 2021), audio (Manor and Michaeli, 2024), and other types of data(Perera et al., 2023; Luo and Hu, 2021; Yang et al., 2023).

Smith and Geach (2019) present a novel approach for simulating galaxy images using DDPMs. DDPMs are employed to generate realistic mock images of galaxies, aiming to mimic observations from the Dark Energy Spectroscopic Instrument (DESI) (Levi et al., 2019) and the Sloan Digital Sky Survey Kollmeier et al. (2019). The synthetic galaxies produced by this method are shown to be highly realistic and comparable to real datasets.

Following this work, a DDPM was trained on the light curves of periodic variable stars from the PRIMVS catalogue. The primary motivation for this is the creation of large training sets. The training set showing in chapter 2 is created partially from synthetic light curves generated using a toy model of trigonometric functions and an attempt at modelling VVV noise and sampling. The method shown in chapter 2 would benefit from a larger and more realistic data set.

The network underwent training for approximately 3 months utilizing 3 NVIDIA Tesla V100 GPUs. It incorporates four channels in its training: magnitude, magnitude error, and the time axis of the light curve normalised by its period, separated into integer and decimal components. Figure 5.5 shows the loss from the model during training. It can be seen that the first ~ 100 steps show little decrease in the loss followed by a large decrease. The trend at the end of figure 5.5

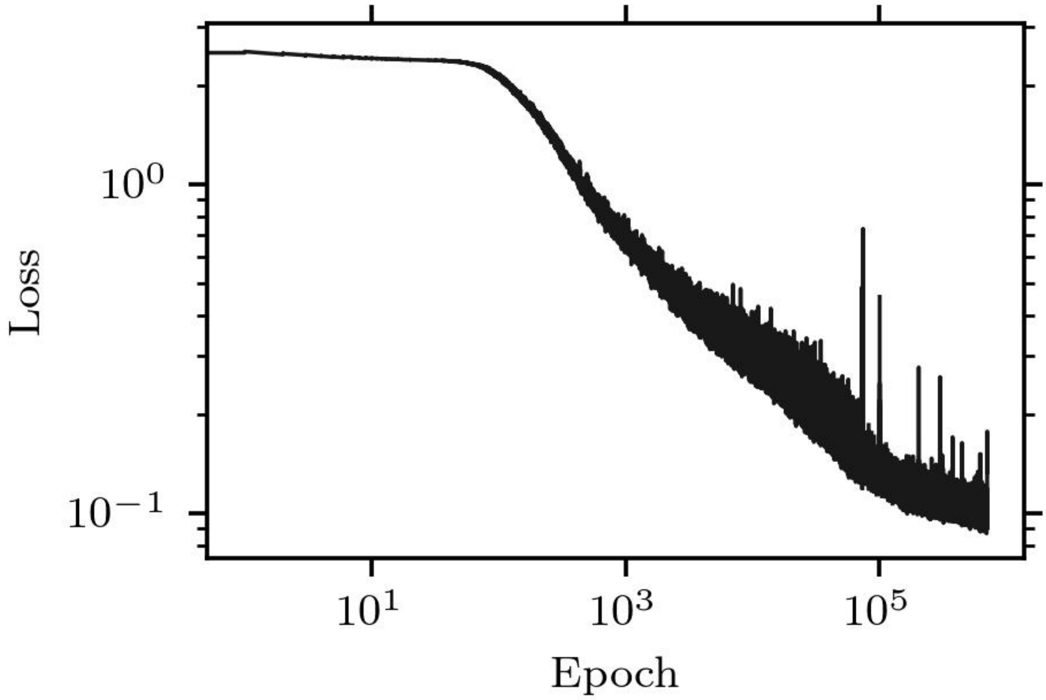


FIGURE 5.5: Showing the loss from training the DDPM.

indicates that training could still continue to achieve a lower loss. The training was stopped due to the large amount of compute time used. A future version will be trained exhaustively.

Figure 5.6 shows the fake light curves generated by this method. It can be seen that apparently realistic periodic variable light curves are generated. This is encouraging as both the time and phase axis seem realistic. These results are promising as the network was trained in a way that requires the periodic signal to be encoded into the time axis.

Future work for this method will be aimed towards obtaining a larger, pre-labelled dataset.

Extensive statistical analysis is required to verify the realistic nature of these lightcurves. A future version would also seek to train with respect to stellar class such that it would be possible to generate periodic variable lightcurves of a desired stellar class. This would be key to improving the work presented in chapter 1.

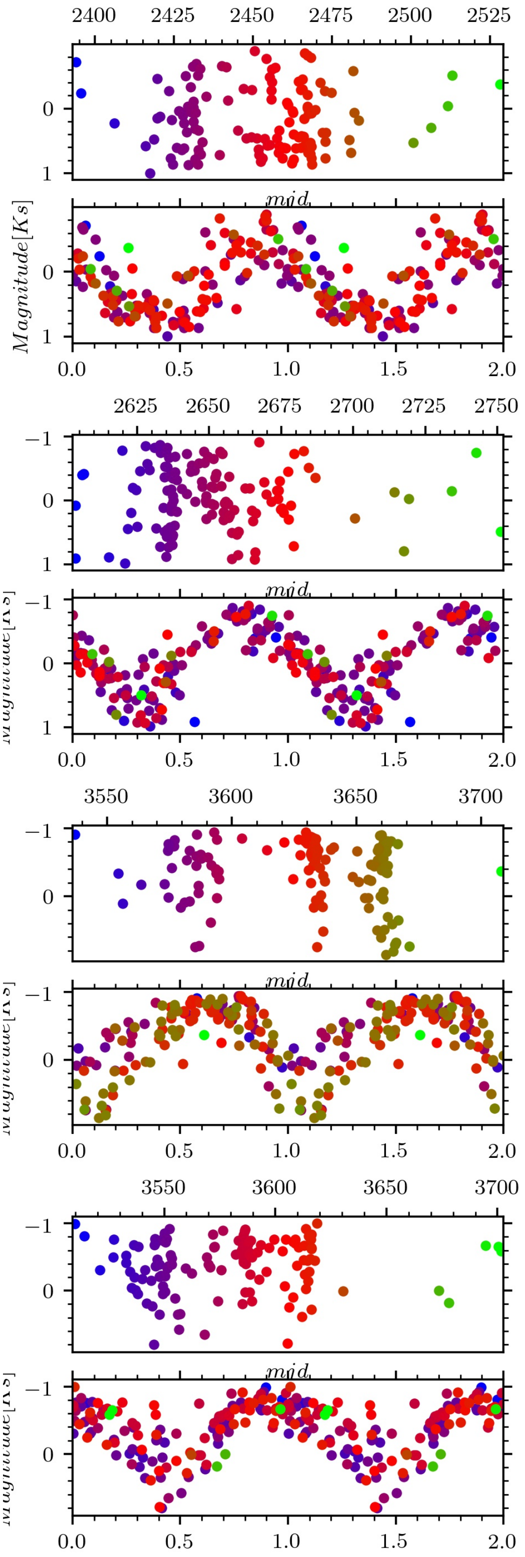


FIGURE 5.6: Fake light curves generated with the trained DDPM. Top: the light curve as a function of time. Bottom: the phase-folded light curve

5.3 Research Proposal

The following is my research proposal for post-doctoral work. I have accepted an offer at the University of Wyoming with this research proposal forming some of my initial application for the position.

Context: Large Astronomical data and its challenges

The last decade in astronomy has seen the growth of time-series data and with it, the emergence of large surveys. Surveys such as PTF (Law et al., 2009), ZTF (Bellm et al., 2019), CoRoT (Auvergne et al., 2009), HOYS (Froebrich et al., 2018) and VVV (Minniti et al., 2010) provide large amounts of large-area, multi-epoch data. Such surveys bring a multitude of new issues, many of which are in the form of 'unknown-unknowns'. From this, novel techniques are required to properly analyse these data. Manual analysis is unfeasible and hence, efforts have been taken to develop tools that seek to automate large portions of the data analysis. The new dimension of study afforded to us by these surveys allows us to probe the formation, evolution and death of stars in unique ways.

A fundamental issue arises **"How can we completely and robustly extract information from modern astronomical time series data?"** – From this question, I aim to further explain the demographics of variable stars and characterise new classes. This seeks to provide a more complete and accurate view of the Milky Way.

An example of this is in the search for periodic variable stars. The 'traditional' Baluev False Alarm Probability (FAP) (Baluev, 2008) features a bias as a function of the shape of the light curve, as seen in Figure 5.7 (top). The Baluev FAP, and by extension, the Lomb-Scargle (Lomb, 1976; Scargle, 1982) periodogram, is fundamentally creating a measure of how sinusoidal a light curve is. Hence, any analysis of periodic star catalogues produced with this method will also feature such a bias. I have produced a recurrent neural network-based method which sought to remove such a bias by fundamentally changing the nature of how we verify periodicity (submitted). If we instead identify only 'structure' in a phase folded light curve then we can get a universal measure of periodicity from a relatively simplistic and easy-to-understand neural network.

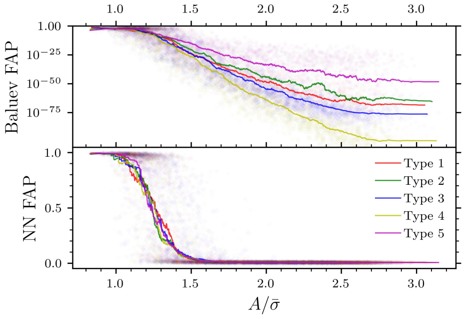


FIGURE 5.7: Showing the False Alarm Probability for synthetic periodic variable stars as a function of the signal-to-noise ratio. Each colour represents a type of synthetic light curve (ranging from sinusoidal to eclipsing binary-like). **[Top]**The Baluev False Alarm Probability shows a bias is imparted with respect to the shape of the light curve. **[Bottom]** The neural network False Alarm Probability

Machine learning, and its principles, can help more completely analyse such unwieldy data. The significantly more flexible analysis methods provided by deep learning offer greater expandability and more room to conserve information from raw data. Typically, these advances are first seen in the extragalactic community. This is largely due to heuristic measures already being difficult when dealing with Galactic morphology and so a greater incentive to produce such methods is seen (Domínguez Sánchez et al., 2018; Gharat and Dandawate, 2022; Martin et al., 2020). The stellar community has been slower in adopting these and time series stellar astronomy is primed for such tools. This reluctance to these more modern methods can be seen in a review of stellar classification. Historically, stellar classification is performed by the heuristic identification of features and a comparison of those features to (typically observationally) defined templates (Robitaille, 2008; Knigge et al., 2011; Koenig and Leisawitz, 2014).

Optical variability, spanning from a few tenths to several magnitudes in brightness, characterises Young Stellar Objects (YSOs). Previous studies, such as Joy (1945), Herbig (1954), and Cody and Hillenbrand (2018), have relied on specific criteria to identify these YSOs. Yet, these classifications are limited by their rigidly defined heuristic features. A more data-centric approach could offer deeper insights into YSO classifications and enhance our understanding of these and similar stellar object classes.

More robust treatment of stellar variability allows for probing into unique areas of astrophysics otherwise unavailable. Key amongst these areas is the star-disk interactions. It is possible for instruments like ALMA to probe protoplanatery disks to a certain level (van Terwisga et al., 2022) but the specifics of the inner disk interactions are best and more easily studied by photometric variability. We know from just looking into periodic accretion variability that the diversity present far exceeds classical modelling and techniques (Audard et al., 2014), this constitutes just one facet of a greater whole when referring to YSO variability.

Taking inspiration from Sarmiento et al. (2021) I have developed a method to self-supervise the classification of variable stars via their light curves (in prep). This, in conjunction with external features (such as spectra), creates a much less rigid base for classifying stars. Furthermore, this method provides a representation of astronomical light curves without the need for a prior. Previously, it was difficult to find more of an exotic object if only 1, or a few, were known. How could a human properly constrain the measured features of a star to define its class? With this method, we need to only identify nearest neighbours in the latent space representation of the data. This method is currently in its infancy and at the time of writing, SimCLR (Chen et al., 2020) is 3 years old and will be superseded (Brown et al., 2020).

The *Gaia* mission (Gaia Collaboration et al., 2016) provides precise astrometric measurements of more than a billion stars giving us an unprecedented 3D view of the stars (and their motions) in our neighbourhood of the Galaxy. However, full exploitation of this dataset for star formation requires combination with other multiwavelength survey datasets and extensive work to identify young stars, derive their properties, and characterise their spatial distributions and dynamics. I have used Multilayer Perceptrons to supplement Gaia astrometry with VVV astrometry. VVV is a multi-epoch IR survey of the southern Galactic disk and bulge, its astrometric measurements compare to Gaia in terms of resolution but not precision. However, in regions of high optical obscuration, such as star-forming regions, Gaia can quickly begin to fail. This is of particular issue when differentiating between background giant stars and foreground Young Stellar Objects (YSOs). The combination of these two surveys provides a similar accuracy to Gaia with the completeness of VVV.

With my research, I aim to identify and solve sources of bias and lost information in time series (and more broadly, astronomical) data sets. From this, I aim produce a more complete and unbiased view of the stellar demographics of the Milky Way by acknowledging and tackling these weaknesses.

5.4 Objectives

My objectives include 1) Identifying and tackling sources of bias and loss of information in time-series (and broadly, astronomical) data sets by developing and using novel techniques 2) improving our understanding of stellar demographics, leading to a better picture of Milky Way structure, and 3) using machine learning tools to more appropriately combine data across multiple surveys and catalogues.

5.5 Methodology

(Periodic) Variable Star Identification and Analysis:

I have developed statistical and machine learning-based methods for identifying and characterising variable stars. I have used these tools to construct PRIMVS (in prep), the most complete catalogue for periodic IR variable stars. These tools can be further improved and applied to other data sets. I will investigate their improvement by expanding on the PRIMVS pipeline. One area of improvement can be made with variability characterised with the use of Gaussian Processes and potentially the more generalisable Deep Gaussian Processes (Damianou and Lawrence, 2012).

Furthermore, much more can be done in the characterisation and classification of variable stars. Self-supervised learning is at the forefront of such open-ended problems and an exploration of natural language processing methods is a necessary future for astronomical light curves (i.e. the creation of a ‘variable star-GPT’)

I will classify and characterise every variable star in the VVV dataset ($\approx 10,000,000$). I will use the pipeline created for this on other data sets to improve on previous catalogues, classifications and characterisations. This will not only allow for a higher completeness but also a more robust and comparable study of stellar variability.

Combining Data:

I have developed a method to combine VVV parallaxes with Gaia parallaxes. This has proven to be incredibly useful for areas looking into the Galactic disk, such as star-forming regions. An obvious next step with this work is to add more surveys, such as the upcoming LSST. This approach of aggregating surveys via neural networks is also applicable to other physical features such as proper motion and SED construction.

We can also combine catalogues instead of surveys. Figure 5.8 shows candidate new class II YSOs. These candidates were identified by cross-matching the VVV-based PRIMVS catalogue to the Spitzer/IRAC-based SPICY (Kuhn et al., 2021). This cross-matched data was used as a training set to find similar objects in the PRIMVS catalogue with the use of an auto-encoder. We can effectively use the careful selections made to construct the SPICY catalogue to parameterise a selection YSOs in the PRIMVS catalogue.

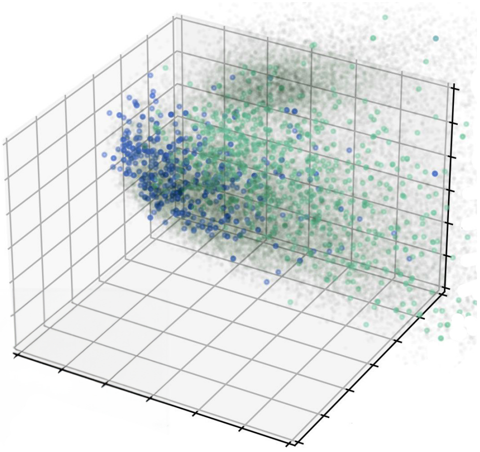


FIGURE 5.8: Showing the latent space (PCA) representation of known SPICY candidate class II YSOs (blue) and PRIMVS candidate class II YSOs (green).

Chapter 6

Conclusions

This thesis explores the application of modern data science techniques to the study of variable stars within the Milky Way, showcasing significant advancements in methodology and astronomical understanding.

Chapter 2 focuses on the development and implementation of Recurrent Neural Networks (RNNs) to process time-series data from the VVV survey. The goal was to verify periodicity in phase-folded light curves, providing a reliable and unbiased method for detecting periodic variable stars. Traditional methods for determining false alarm probabilities (FAP) often fall short due to assumptions of light curve shape and Gaussian errors. We addressed these limitations by presenting a novel machine learning-based technique that directly analyses phase-folded light curves. The RNNs employed in this study were trained on both real and synthetic light curves of periodic and aperiodic sources. This approach allowed the model to be trained on a large range of light curve shapes, improving its ability to generalise across different types of variable stars. By carefully controlling the types of light curves presented and employing various data augmentation and normalisation techniques, dependencies on specific light curve shapes are minimised. Our approach demonstrated an improvement in detecting periodic variables, showing even recovery rates across multiple light curve shapes. The results also highlighted the model's ability to differentiate between periodic and aperiodic light curves with varying numbers of data points and signal-to-noise ratios. Further to this, the practical applications of this method was shown in large-scale astronomical surveys like OGLE, ZTF, Kepler, and TESS, where the volume of data makes traditional analysis methods impractical.

The neural network-based FAP (NN FAP) offers an universal, scalable measure of periodicity that is largely insensitive to the specific shapes of light curves, making it applicable across a diverse array of variable star types without extensive preprocessing or model-specific tuning. By automating the detection and parameterisation of variable stars, this method enhances the efficiency of data processing, making it feasible to analyse the vast datasets generated by modern astronomical surveys. The NN FAP method's insensitivity to light curve shapes allows it to be applied across various types of variable stars, including eclipsing binaries and pulsating stars. This versatility makes it a valuable tool for characterising the diverse population of variable stars in the Milky Way and beyond. Additionally, the ability to handle sparse and noisy data ensures that even less well-sampled surveys can benefit from this approach, broadening the scope of astronomical research. Future research can build on this foundation by integrating more sophisticated machine learning models and exploring their applications in other areas of astronomy. For instance, the development of autoencoders and diffusion models for generating synthetic light curves can further enhance the model's capabilities.

The PeRiodic Infrared Milky-way VVV Star-catalogue (PRIMVS) represents a significant advancement in the study of variable stars within the Galactic bulge and disk, utilising the unique depth and breadth of the VVV survey coupled with modern data science techniques. The PRIMVS pipeline cleans and preprocesses light curves by addressing photometric contamination and other uncertainties using metrics provided by VIRAC. Multiple period-finding methods such as Lomb-Scargle, Phase Dispersion Minimisation, Conditional Entropy, and Gaussian Processes are used and their calculated periodicities compared on a uniform scale using the NN FAP. This has enabled the identification and classification of over 86 million candidate variable sources, including approximately 5 million periodic variables. The integration of machine learning techniques, particularly decision trees, facilitated the classification of a substantial portion of these sources. Cross-matched data from Gaia DR3 provided a robust training set, although this approach introduced potential biases due to differences in depth and selection criteria between the Gaia and VVV surveys. The decision tree model achieved high classification accuracy, successfully identifying known classes of stars and revealing insights into the stellar population of the Milky Way. The use of autoencoders and UMAP for latent space representation furthered the analysis, highlighting distinct groups of quasi-periodic sources. PRIMVS has demonstrated the effectiveness of combining traditional astronomical analysis with modern data science techniques. The catalogue's thorough preprocessing and the application of advanced period-finding methods have resulted in a comprehensive and reliable dataset, providing a reliable source for future studies of variable stars. The use of a novel FAP method to enhance the reliability of period identification is a notable contribution. The PRIMVS pipeline demonstrates the potential for further advancements in astronomical data analysis. Applying the PRIMVS pipeline to other large-scale time-domain surveys, such as LSST and TESS, could provide a more comprehensive understanding of variable stars across different regions of the sky and wavelengths.

The use of contrastive learning to analyse light curves is shown in chapter 3.8, marking a notable step in handling time-series data in astronomy. This approach centres around the SimCLR architecture with a gated recurrent neural network (GRU) backbone, specifically designed to process stochastically sampled time-series data. This architecture not only addresses the inherent challenges of irregular sampling but also ensures that the extracted representations are semantically meaningful. A series of data augmentation techniques are used that are crucial for contrastive learning. These augmentations modify the light curves in realistic ways that do not fundamentally change the represented stellar class. This allows for the creation of semantically useful embeddings. These embeddings are essential as they encapsulate the intrinsic properties of the light curves, allowing for more nuanced and open-ended analysis. By minimising the normalised temperature cross-entropy (NT-Xent) loss, similar light curves are clustered together in the embedding space, while dissimilar ones are pushed apart. This clustering is pivotal in identifying and classifying variable stars with high accuracy. The latent space representations generated by this model were validated against known astronomical datasets. This approach not only improves the accuracy and reliability of variable star classification but also allows for future applications in large-scale surveys like LSST and TESS.

The integration of VVV data with Gaia astrometry can enhance the accuracy of distance measurements to star-forming regions, often obscured by interstellar dust. Gaia provides precise astrometric measurements but struggles in dusty regions where optical signals are scattered and absorbed. VVV, operating in the near-infrared, can penetrate these dusty regions, offering complementary data to Gaia. By employing a Multilayer Perceptron (MLP) to merge VVV and Gaia data, an improved parallax can be generated, particularly in areas where Gaia's optical observations are hindered by dust. This combined approach helps disentangle Young Stellar Objects (YSOs) from their contaminating Asymptotic Giant Branch (AGB) counterparts, providing a clearer picture of star-forming regions. Initial results indicate that this method effectively matches Gaia distances for stars near the Galactic Center, though challenges remain for more distant sources.

Denoising Diffusion Probabilistic Models (DDPMs) offer a way of generating synthetic light curves, addressing the need for large, realistic training sets in machine learning applications. Trained on the light curves of periodic variable stars from the PRIMVS catalogue, DDPMs can produce high-quality, diverse samples that closely resemble real astronomical data. This capability is helpful in developing robust machine learning models that require extensive datasets to achieve high accuracy (such as what is shown in chapter **??**. Future work aims to refine the training process, potentially incorporating specific stellar classes to generate desired light curves.

This thesis presents a comprehensive investigation into the application of modern data science techniques to the study of variable stars within the Milky Way. This work spans the development of innovative machine learning algorithms, the creation of an extensive star catalogue, and its novel analysis.

The first major contribution of this thesis is the development and implementation of Recurrent Neural Networks (RNNs) to process time-series data from astronomical surveys. By addressing the limitations of traditional false alarm probability (FAP) methods, the thesis introduces a neural network-based FAP (NN FAP) that significantly improves the detection and characterisation of periodic variables. This method, demonstrated on data from the VISTA Variables in the Vía Láctea (VVV) survey, shows great potential for application in other large-scale astronomical surveys like LSST, ZTF, Kepler, and TESS, thereby enhancing the efficiency and reliability of variable star analysis on a broad scale.

The second significant contribution is the creation of the PeRiodic Infrared Milky-way VVV Star-catalogue (PRIMVS) By utilising the depth and breadth of the VVV survey and employing sophisticated data preprocessing and period-finding techniques, the PRIMVS catalogue has identified and classified over 86 million candidate variable sources, including approximately 5 million periodic variables. This extensive dataset provides a valuable resource for future astronomical studies, offering new insights into the distribution and characteristics of variable stars within our galaxy.

The third major advancement highlighted in this thesis is the application of contrastive learning methods to the analysis of light curves. Utilising a novel neural network architecture based on the SimCLR framework with a gated recurrent neural network (GRU) backbone, this approach effectively handles the complexities of stochastically sampled time-series data. The resulting semantically meaningful embeddings enable more accurate and nuanced identification characterisation and, classification of variable stars.

The thesis also explores the integration of data from the VVV survey with Gaia astrometry to improve distance measurements to star-forming regions. This combined approach addresses the limitations of optical observations in dusty regions, offering clearer insights into the distribution and characteristics of Young Stellar Objects (YSOs). Additionally, the exploration of Denoising Diffusion Probabilistic Models (DDPMs) for generating synthetic light curves demonstrates a promising avenue for creating realistic training sets.

This thesis highlights the potential of integrating traditional astronomical analysis with modern and novel data science techniques. The methodologies developed and applied in this work not only enhance the efficiency and accuracy of variable star detection and classification but also provide powerful tools for future unconsidered astronomical research. The findings and techniques presented here will continue to influence and drive advancements in the field of astronomy.

Bibliography

1950. Hebb, d. o. the organization of behavior: A neuropsychological theory. new york: John wiley and sons, inc., 1949. 335 p. $4.00. *Science Education*, 34(5):336.

2023. *Protostars and Planets VII*, volume 534 of *Astronomical Society of the Pacific Conference Series*.

Abul-Hayat, M., Stein, G., Harrington, P., et al., 2020. Self-Supervised Representation Learning for Astronomical Images. *arXiv e-prints*, arXiv:2012.13083.

Anderson, T.W. and Darling, D.A., 1952. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193 .

Angus, R., Morton, T., Aigrain, S., et al., 2018. Inferring probabilistic stellar rotation periods using Gaussian processes. *MNRAS*, 474(2):2094.

Astropy Collaboration, Robitaille, T.P., Tollerud, E.J., et al., 2013. Astropy: A community Python package for astronomy. *A&A*, 558:A33.

Audard, M., Ábrahám, P., Dunham, M.M., et al., 2014. Episodic Accretion in Young Stars. In H. Beuther, R.S. Klessen, C.P. Dullemond, and T. Henning, editors, *Protostars and Planets VI*, pages 387–410.

Auvergne, M., Bodin, P., Boisnard, L., et al., 2009. The CoRoT satellite in flight: description and performance. *A&A*, 506(1):411.

BAART, M.L., 1982. The Use of Auto-correlation for Pseudo-rank Determination in Noisy III-conditioned Linear Least-squares Problems. *IMA Journal of Numerical Analysis*, 2(2):241.

Bailey, S.I., Leland, E.F., Woods, I.E., et al., 1919. Variable stars in the cluster Messier 15. *Annals of Harvard College Observatory*, 78:195.

Baluev, R.V., 2008. Assessing the statistical significance of periodogram peaks. *MNRAS*, 385(3):1279.

Baluev, R.V., 2009. Detecting non-sinusoidal periodicities in observational data using multiharmonic periodograms. *MNRAS*, 395(3):1541.

Becker, I., Pichara, K., Catelan, M., et al., 2020. Scalable end-to-end recurrent neural network for variable star classification. *MNRAS*, 493(2):2981.

Bell, K.R., Lin, D.N.C., Hartmann, L.W., et al., 1995. The FU Orionis Outburst as a Thermal Accretion Event: Observational Constraints for Protostellar Disk Models. *ApJ*, 444:376.

Bellm, E.C., Kulkarni, S.R., Graham, M.J., et al., 2019. The Zwicky Transient Facility: System Overview, Performance, and First Results. *PASP*, 131(995):018002.

Belokurov, V., Evans, N.W., and Du, Y.L., 2003. Light-curve classification in massive variability surveys - I. Microlensing. *Monthly Notices of the Royal Astronomical Society*, 341(4):1373.

Bhardwaj, A., Panwar, N., Herczeg, G.J., et al., 2019. Variability of young stellar objects in the star-forming region Pelican Nebula. *A&A*, 627:A135.

Bobrick, A., Iorio, G., Belokurov, V., et al., 2024. RR Lyrae from binary evolution: abundant, young, and metal-rich. *MNRAS*, 527(4):12196.

Bono, G., Marconi, M., and Stellingwerf, R.F., 2000. Classical Cepheid pulsation models — VI. The Hertzsprung progression. *A&A*, 360:245.

Bornfreund, R.E., 2005. Large format short-wave hgcdte detectors and focal plane arrays. *Raytheon Vision Systems*.

Borucki, W.J., Koch, D., Basri, G., et al., 2003. Kepler Mission: a mission to find Earth-size planets in the habitable zone. In M. Fridlund, T. Henning, and H. Lacoste, editors, *Earths: DARWIN/TPF and the Search for Extrasolar Terrestrial Planets*, volume 539 of *ESA Special Publication*, pages 69–81.

Botan, E., Saito, R.K., Minniti, D., et al., 2021. Unveiling short-period binaries in the inner VVV bulge. *MNRAS*, 504(1):654.

Bountos, N.I., Papoutsis, I., Michail, D., et al., 2022. Self-Supervised Contrastive Learning for Volcanic Unrest Detection. *IEEE Geoscience and Remote Sensing Letters*, 19:3104506.

Braga, V.F., Bono, G., Fiorentino, G., et al., 2020. Separation between RR Lyrae and type II Cepheids and their importance for a distance determination: the case of omega Cen. *A&A*, 644:A95.

Brett, D.R., West, R.G., and Wheatley, P.J., 2004. The automated classification of astronomical light curves using Kohonen self-organizing maps. *MNRAS*, 353(2):369.

Brodrick, D., Taylor, D., and Diederich, J., 2004. Recurrent neural networks for narrowband signal detection in the time-frequency domain. *Symposium - International Astronomical Union*, 213:483–486.

Brown, T.B., Mann, B., Ryder, N., et al., 2020. Language Models are Few-Shot Learners. *arXiv e-prints*, arXiv:2005.14165.

Bundy, K., Bershady, M.A., Law, D.R., et al., 2015. Overview of the SDSS-IV MaNGA Survey: Mapping nearby Galaxies at Apache Point Observatory. *ApJ*, 798(1):7.

Burhanudin, U.F., Maund, J.R., Killestein, T., et al., 2021. Light-curve classification with recurrent neural networks for GOTO: dealing with imbalanced data. *MNRAS*, 505(3):4345.

Cabrera Garcia, J., Beers, T.C., Huang, Y., et al., 2023. Probing the Galactic halo with RR Lyrae stars – V. Chemistry, kinematics, and dynamically tagged groups. *Monthly Notices of the Royal Astronomical Society*, 527(3):8973.

Capizzi, G., Napoli, C., and Paternò, L., 2012. An innovative hybrid neuro-wavelet method for reconstruction of missing data in astronomical photometric surveys. In L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L.A. Zadeh, and J.M. Zurada, editors, *Artificial Intelligence and Soft Computing*, pages 21–29. Springer Berlin Heidelberg, Berlin, Heidelberg.

Carrasco-Davis, R., Cabrera-Vives, G., Förster, F., et al., 2019. Deep learning for image sequence classification of astronomical events. *Publications of the Astronomical Society of the Pacific*, 131(1004):108006.

Catchpole, R.M., Whitelock, P.A., Feast, M.W., et al., 2016. The age and structure of the Galactic bulge from Mira variables. *MNRAS*, 455(2):2216.

Charnock, T. and Moss, A., 2017. Deep Recurrent Neural Networks for Supernovae Classification. *The Astrophysical Journal Letters*, 837(2):L28.

Chen, T. and Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. *arXiv e-prints*, arXiv:1603.02754.

Chen, T., Kornblith, S., Norouzi, M., et al., 2020. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv e-prints*, arXiv:2002.05709.

Chen, T., Kornblith, S., Swersky, K., et al., 2020a. Big self-supervised models are strong semi-supervised learners. *CoRR*, abs/2006.10029.

Chen, X., Fan, H., Girshick, R.B., et al., 2020b. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297.

Chen, X., Wang, S., Deng, L., et al., 2019. An intuitive 3d map of the galactic warp's precession traced by classical cepheids. *Nature Astronomy*, 3(4):320.

Chen, X., Wang, S., Deng, L., et al., 2020. The Zwicky Transient Facility Catalog of Periodic Variable Stars. *ApJ*, 249(1):18.

Chen, X., Zhang, J., Wang, S., et al., 2023. The use of double-mode rr lyrae stars as robust distance and metallicity indicators. *Nature Astronomy*, 7(9):1081.

Chibueze, J.O., Miyahara, T., Omodaka, T., et al., 2016. Near-infrared Observations of SiO Maser-emitting Asymptotic Giant Branch (AGB) Stars. *ApJ*, 817(2):115.

Cho, K., van Merrienboer, B., Bahdanau, D., et al., 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

Cho, K., van Merrienboer, B., Gulcehre, C., et al., 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv e-prints*, arXiv:1406.1078.

Chollet, F. et al., 2015. Keras. `https://keras.io`.

Christy, C.T., Jayasinghe, T., Stanek, K.Z., et al., 2023. The ASAS-SN catalogue of variable stars X: discovery of 116 000 new variable stars using G-band photometry. *MNRAS*, 519(4):5271.

Christy, R.F., 1975. The Hertzsprung progression in Cepheid calculations. In *NASA Special Publication*, volume 383, pages 85–98.

Cincotta, P.M., Mendez, M., and Nunez, J.A., 1995. Astronomical Time Series Analysis. I. A Search for Periodicity Using Information Entropy. *ApJ*, 449:231.

Clayton, G.C., 1996. The R Coronae Borealis Stars. *PASP*, 108:225.

Cody, A.M. and Hillenbrand, L.A., 2018. The Many-faceted Light Curves of Young Disk-bearing Stars in Upper Sco – Oph Observed by K2 Campaign 2. *AJ*, 156(2):71.

Cody, A.M., Stauffer, J., Baglin, A., et al., 2014. CSI 2264: Simultaneous Optical and Infrared Light Curves of Young Disk-bearing Stars in NGC 2264 with CoRoT and Spitzer—Evidence for Multiple Origins of Variability. *AJ*, 147(4):82.

Contreras Peña, C., Lucas, P.W., Froebrich, D., et al., 2014. Extreme infrared variables from UKIDSS - I. A concentration in star-forming regions. *MNRAS*, 439(2):1829.

Contreras Peña, C., Lucas, P.W., Minniti, D., et al., 2017. A population of eruptive variable protostars in VVV. *MNRAS*, 465(3):3011.

Corradi, R.L.M., Rodríguez-Flores, E.R., Mampaso, A., et al., 2008. IPHAS and the symbiotic stars. I. Selection method and first discoveries. *A&A*, 480(2):409.

Covey, K.R., Larson, K.A., Herczeg, G.J., et al., 2021. A Differential Measurement of Circumstellar Extinction for AA Tau's 2011 Dimming Event. *AJ*, 161(2):61.

Cox, J.P., 1963. On Second Helium Ionization as a Cause of Pulsational Instability in Stars. *ApJ*, 138:487.

da Silva, R., Crestani, J., Bono, G., et al., 2022. A new and Homogeneous metallicity scale for Galactic classical Cepheids. II. Abundance of iron and α elements. *A&A*, 661:A104.

Damianou, A.C. and Lawrence, N.D., 2012. Deep Gaussian Processes. *arXiv e-prints*, arXiv:1211.0358.

Deb, S. and Singh, H.P., 2009. Light curve analysis of variable stars using Fourier decomposition and principal component analysis. *A&A*, 507(3):1729.

Deeming, T.J., 1975. Fourier Analysis with Unequally-Spaced Data. , 36(1):137.

Devlin, J., Chang, M.W., Lee, K., et al., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota.

Dewdney, P.E., Hall, P.J., Schilizzi, R.T., et al., 2009. The Square Kilometre Array. *IEEE Proceedings*, 97(8):1482.

Dieleman, S., Willett, K.W., and Dambre, J., 2015. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *MNRAS*, 450(2):1441.

Distefano, E., Lanzafame, A.C., Lanza, A.F., et al., 2012. Determination of rotation periods in solar-like stars with irregular sampling: the Gaia case. *MNRAS*, 421(4):2774.

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., et al., 2018. Improving galaxy morphologies for SDSS with Deep Learning. *MNRAS*, 476(3):3661.

Donoso-Oliva, C., Becker, I., Protopapas, P., et al., 2022. ASTROMER: A transformer-based embedding for the representation of light curves. *arXiv e-prints*, arXiv:2205.01677.

Drake, A.J., Djorgovski, S.G., Mahabal, A., et al., 2009. First Results from the Catalina Real-Time Transient Survey. *ApJ*, 696(1):870.

Dubath, P., Rimoldini, L., Süveges, M., et al., 2011. Random forest automated supervised classification of Hipparcos periodic variable stars. *MNRAS*, 414(3):2602.

Duchêne, G. and Kraus, A., 2013. Stellar multiplicity. *Annual Review of Astronomy and Astrophysics*, 51(1):269.

Dworetsky, M.M., 1983. A period-finding method for sparse randomly spaced observations or “How long is a piece of string ?”. *MNRAS*, 203:917.

Eddington, A.S., 1988. *The Internal Constitution of the Stars*. Cambridge Science Classics. Cambridge University Press.

Eggen, O.J., 1948. The System of Algol. *ApJ*, 108:1.

Ellaway, P., 1978. Cumulative sum technique and its application to the analysis of peristimulus time histograms. *Electroencephalography and Clinical Neurophysiology*, 45(2):302.

Emmanoulopoulos, D., McHardy, I.M., and Papadakis, I.E., 2013. Generating artificial light curves: revisited and updated. *Monthly Notices of the Royal Astronomical Society*, 433(2):907.

Eyer, L. and Blake, C., 2005. Automated classification of variable stars for All-Sky Automated Survey 1-2 data. *MNRAS*, 358(1):30.

Eyer, L. and Mowlavi, N., 2008. Variable stars across the observational HR diagram. In *Journal of Physics Conference Series*, volume 118 of *Journal of Physics Conference Series*, page 012010.

Fernley, J.A., Skillen, I., Jameson, R.F., et al., 1990. The absolute magnitudes of RR Lyrae stars - III. DH Peg. *MNRAS*, 242:685.

Finke, T., Krämer, M., and Manconi, S., 2021. Classification of Fermi-LAT sources with deep learning using energy and time spectra. *Monthly Notices of the Royal Astronomical Society*, 507(3):4061.

Fix, E. and Hodges, J.L., 1951. Discriminatory analysis: Nonparametric discrimination: Consistency properties. *PsycEXTRA Dataset*.

Froebrich, D., Campbell-White, J., Scholz, A., et al., 2018. A survey for variable young stars with small telescopes: First results from HOYS-CAPS. *MNRAS*, 478(4):5091.

F.R.S., K.P., 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559.

Gaia Collaboration, Brown, A.G.A., Vallenari, A., et al., 2018. Gaia Data Release 2. Summary of the contents and survey properties. *A&A*, 616:A1.

Gaia Collaboration, Brown, A.G.A., Vallenari, A., et al., 2021. Gaia Early Data Release 3. Summary of the contents and survey properties. *A&A*, 649:A1.

Gaia Collaboration, Prusti, T., de Bruijne, J.H.J., et al., 2016. The Gaia mission. *A&A*, 595:A1.

Gaia Collaboration, Vallenari, A., Brown, A.G.A., et al., 2023. Gaia Data Release 3. Summary of the content and survey properties. *A&A*, 674:A1.

Gharat, S. and Dandawate, Y., 2022. Galaxy classification: a deep learning approach for classifying Sloan Digital Sky Survey images. *MNRAS*, 511(4):5120.

Gillet, D., 2013. Atmospheric dynamics in RR Lyrae stars. The Blazhko effect. *A&A*, 554:A46.

Gliozzi, M., Brinkmann, W., Räth, C., et al., 2002. On the nature of X-ray variability in Ark 564. *A&A*, 391:875.

Gomez Gonzalez, C.A., Absil, O., and Van Droogenbroeck, M., 2018. Supervised detection of exoplanets in high-contrast imaging sequences. *A&A*, 613:A71.

Gonzalez, A.G., Gallo, L.C., Miller, J.M., et al., 2023. Characterizing X-ray, UV, and optical variability in NGC6814 using high-cadence Swift observations from a 2022 monitoring campaign. *Monthly Notices of the Royal Astronomical Society*, 527(3):5569.

Graham, M.J., Drake, A.J., Djorgovski, S.G., et al., 2013a. A comparison of period finding algorithms. *MNRAS*, 434(4):3423.

Graham, M.J., Drake, A.J., Djorgovski, S.G., et al., 2013b. Using conditional entropy to identify periodicity. *MNRAS*, 434(3):2629.

Grill, J., Strub, F., Altché, F., et al., 2020. Bootstrap your own latent: A new approach to self-supervised learning. *CoRR*, abs/2006.07733.

Gruberbauer, M., Kolenberg, K., Rowe, J.F., et al., 2007. MOST photometry of the RRd-Lyrae variable AQLeo: two radial modes, 32 combination frequencies and beyond. *MNRAS*, 379(4):1498.

Guo, Z., Lucas, P.W., Contreras Peña, C., et al., 2020. Short- and long-term near-infrared spectroscopic variability of eruptive protostars from VVV. *MNRAS*, 492(1):294.

Guo, Z., Lucas, P.W., Contreras Peña, C., et al., 2021. Analysis of physical processes in eruptive YSOs with near-infrared spectra and multiwavelength light curves. *MNRAS*, 504(1):830.

Hajdu, G., Dékány, I., Catelan, M., et al., 2020. On the optimal calibration of VVV photometry. *Experimental Astronomy*, 49(3):217.

Hála, P., 2014. Spectral classification using convolutional neural networks. *arXiv e-prints*, arXiv:1412.8341.

Hayat, M.A., Stein, G., Harrington, P., et al., 2021. Self-supervised representation learning for astronomical images. *The Astrophysical Journal Letters*, 911(2):L33.

He, K., Fan, H., Wu, Y., et al., 2019. Momentum contrast for unsupervised visual representation learning. *CoRR*, abs/1911.05722.

He, S., Lin, Z., Yuan, W., et al., 2021. Simultaneous inference of periods and period-luminosity relations for Mira variable stars. *Annals of Applied Statistics*, 15(2):662.

Heck, A., Manfroid, J., and Mersch, G., 1985. On period determination methods. *A&A Supp.*, 59:63.

Herbig, G.H., 1954. Emission-Line Stars Associated with the Nebulous Cluster NGC 2264. *ApJ*, 119:483.

Herbig, G.H., 1966. On the interpretation of FU orionis. *Vistas in Astronomy*, 8(1):109.

Herbst, W., Rhode, K.L., Hillenbrand, L.A., et al., 2000. Rotation in the Orion Nebula Cluster. *AJ*, 119(1):261.

Hertzsprung, E., 1926. On the relation between period and form of the light-curve of variable stars of the δ Cephei type. , 3:115.

Hinton, G.E., Srivastava, N., Krizhevsky, A., et al., 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv e-prints*, arXiv:1207.0580.

Ho, J., Jain, A., and Abbeel, P., 2020. Denoising Diffusion Probabilistic Models. *arXiv e-prints*, arXiv:2006.11239.

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural computation*, 9(8):1735.

Höfner, S. and Olofsson, H., 2018. Mass loss of stars on the asymptotic giant branch. Mechanisms, models and measurements. , 26(1):1.

Hogg, D.W., 2008. Data analysis recipes: Choosing the binning for a histogram. *arXiv e-prints*, arXiv:0807.4820.

Huijse, P., Estevez, P.A., Protopapas, P., et al., 2012. An Information Theoretic Algorithm for Finding Periodicities in Stellar Light Curves. *IEEE Transactions on Signal Processing*, 60(10):5135.

Husseiniova, A., McGill, P., Smith, L.C., et al., 2021. A microlensing search of 700 million VVV light curves. *MNRAS*, 506(2):2482.

Irwin, J., Aigrain, S., Bouvier, J., et al., 2009. The Monitor project: rotation periods of low-mass stars in M50. *MNRAS*, 392(4):1456.

Ivezić, Ž., Kahn, S.M., Tyson, J.A., et al., 2019. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *ApJ*, 873(2):111.

Iwanek, P., Poleski, R., Kozłowski, S., et al., 2023. A Three-dimensional Map of the Milky Way Using 66,000 Mira Variable Stars. *ApJ*, 264(1):20.

Jaeger, H., 2002. Tutorial on training recurrent neural networks, covering bppt, rtrl, ekf and the echo state network approach. *GMD-Forschungszentrum Informationstechnik, 2002.*, 5.

Jayasinghe, T., Kochanek, C.S., Stanek, K.Z., et al., 2018. The ASAS-SN catalogue of variable stars I: The Serendipitous Survey. *MNRAS*, 477(3):3145.

Jayasinghe, T., Stanek, K.Z., Kochanek, C.S., et al., 2019. The ASAS-SN catalogue of variable stars - II. Uniform classification of 412 000 known variables. *MNRAS*, 486(2):1907.

Jiménez-Esteban, F.M., García-Lario, P., Engels, D., et al., 2006. Near-IR variability properties of a selected sample of AGB stars. *A&A*, 458(2):533.

Joy, A.H., 1945. T Tauri Variable Stars. *ApJ*, 102:168.

Joyce, M., Molnár, L., Cinquegrana, G., et al., 2024. Stellar Evolution in Real Time II: R Hydrae and an Open-Source Grid of ¿3000 Seismic TP-AGB Models Computed with MESA. *arXiv e-prints*, arXiv:2401.16142.

Kains, N., Calamida, A., Rejkuba, M., et al., 2019. New variable stars towards the Galactic Bulge - I. The bright regime. *MNRAS*, 482(3):3058.

Karpenka, N.V., Feroz, F., and Hobson, M.P., 2013. A simple and robust method for automated photometric classification of supernovae using neural networks. *MNRAS*, 429(2):1278.

Kesseli, A.Y., Petkova, M.A., Wood, K., et al., 2016. A Model for (Quasi-)Periodic Multiwavelength Photometric Variability in Young Stellar Objects. *ApJ*, 828(1):42.

Kim, D.W. and Bailer-Jones, C.A.L., 2016. A package for the automated classification of periodic variable stars. *A&A*, 587:A18.

Kim, D.W., Protopapas, P., Bailer-Jones, C.A.L., et al., 2014. The EPOCH Project. I. Periodic variable stars in the EROS-2 LMC database. *A&A*, 566:A43.

Kingma, D.P. and Ba, J., 2014. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, arXiv:1412.6980.

Kirk, B., Conroy, K., Prša, A., et al., 2016. Kepler Eclipsing Binary Stars. VII. The Catalog of Eclipsing Binaries Found in the Entire Kepler Data Set. *AJ*, 151(3):68.

Klusch, M. and Napiwotzki, R., 1993. HNS : a hybrid neural system and its use for the classification of stars. *A&A*, 276:309.

Knigge, C., Baraffe, I., and Patterson, J., 2011. The Evolution of Cataclysmic Variables as Revealed by Their Donor Stars. *ApJ*, 194(2):28.

Koeltzsch, A., Mugrauer, M., Raetz, S., et al., 2009. Variability of young stars: Determination of rotational periods of weak-line T Tauri stars in the Cepheus-Cassiopeia star-forming region. *Astronomische Nachrichten*, 330(5):482.

Koenig, X.P. and Leisawitz, D.T., 2014. A Classification Scheme for Young Stellar Objects Using the Wide-field Infrared Survey Explorer AllWISE Catalog: Revealing Low-density Star Formation in the Outer Galaxy. *ApJ*, 791(2):131.

Kollmeier, J., Anderson, S.F., Blanc, G.A., et al., 2019. SDSS-V Pioneering Panoptic Spectroscopy. In *Bulletin of the American Astronomical Society*, volume 51, page 274.

Krizhevsky, A., Sutskever, I., and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Kügler, S.D., Gianniotis, N., and Polsterer, K.L., 2016. An explorative approach for inspecting Kepler data. *MNRAS*, 455(4):4399.

Kuhn, M.A., de Souza, R.S., Krone-Martins, A., et al., 2021. SPICY: The Spitzer/IRAC Candidate YSO Catalog for the Inner Galactic Midplane. *ApJ*, 254(2):33.

Lakeland, B.S. and Naylor, T., 2022. Towards an understanding of YSO variability: a multiwavelength analysis of bursting, dipping, and symmetrically varying light curves of disc-bearing YSOs. *MNRAS*, 514(2):2736.

Lastilla, L., Ammirati, S., Firmani, D., et al., 2022. Self-supervised learning for medieval handwriting identification: A case study from the vatican apostolic library. *Information Processing Management*, 59(3):102875.

Law, N.M., Kulkarni, S.R., Dekany, R.G., et al., 2009. The Palomar Transient Factory: System Overview, Performance, and First Results. *PASP*, 121(886):1395.

Lawrence, A., Warren, S.J., Almaini, O., et al., 2007. The UKIRT Infrared Deep Sky Survey (UKIDSS). *MNRAS*, 379(4):1599.

Leavitt, H.S. and Pickering, E.C., 1912. Periods of 25 Variable Stars in the Small Magellanic Cloud. *Harvard College Observatory Circular*, 173:1.

Levi, M., Allen, L.E., Raichoor, A., et al., 2019. The Dark Energy Spectroscopic Instrument (DESI). In *Bulletin of the American Astronomical Society*, volume 51, page 57.

Lissauer, J.J., Rowe, J.F., Jontof-Hutter, D., et al., 2023. Updated Catalog of Kepler Planet Candidates: Focus on Accuracy and Orbital Periods. *arXiv e-prints*, arXiv:2311.00238.

Liu, G., Huang, Y., Bird, S.A., et al., 2022. Probing the Galactic halo with RR lyrae stars - III. The chemical and kinematic properties of the stellar halo. *MNRAS*, 517(2):2787.

Liu, H., Liu, C., Wang, J.T.L., et al., 2019. Predicting solar flares using a long short-term memory network. *The Astrophysical Journal*, 877(2):121.

Lomb, N.R., 1976. Least-Squares Frequency Analysis of Unequally Spaced Data. , 39(2):447.

Lopez-Vazquez, V., Lopez-Guede, J.M., Chatzievangelou, D., et al., 2023. Deep learning based deep-sea automatic image enhancement and animal species classification. *Journal of Big Data*, 10(1):37.

Lucas, P.W., Hoare, M.G., Longmore, A., et al., 2008. The UKIDSS Galactic Plane Survey. *MNRAS*, 391(1):136.

Lucas, P.W., Smith, L.C., Contreras Peña, C., et al., 2017. Extreme infrared variables from UKIDSS - II. An end-of-survey catalogue of eruptive YSOs and unusual stars. *MNRAS*, 472(3):2990.

Lundberg, S. and Lee, S.I., 2017. A Unified Approach to Interpreting Model Predictions. *arXiv e-prints*, arXiv:1705.07874.

Luo, S. and Hu, W., 2021. Diffusion Probabilistic Models for 3D Point Cloud Generation. *arXiv e-prints*, arXiv:2103.01458.

Maintz, G. and de Boer, K.S., 2005. RR Lyrae stars: kinematics, orbits and z-distribution. *A&A*, 442(1):229.

Mainzer, A., Bauer, J., Cutri, R.M., et al., 2014. Initial Performance of the NEOWISE Reactivation Mission. *ApJ*, 792(1):30.

Manor, H. and Michaeli, T., 2024. Zero-Shot Unsupervised and Text-Based Audio Editing Using DDPM Inversion. *arXiv e-prints*, arXiv:2402.10009.

Martin, G., Kaviraj, S., Hocking, A., et al., 2020. Galaxy morphological classification in deep-wide surveys via unsupervised machine learning. *MNRAS*, 491(1):1408.

Matsunaga, N., Itane, A., Hattori, K., et al., 2022. A Very Metal-poor RR Lyrae Star with a Disk Orbit Found in the Solar Neighborhood. *ApJ*, 925(1):10.

Matsunaga, N., Kawadu, T., Nishiyama, S., et al., 2011. Three classical Cepheid variable stars in the nuclear bulge of the Milky Way. *Nature*, 477(7363):188.

Mattei, J.A., 1997. Introducing Mira Variables. , 25(2):57.

McCulloch, W.S. and Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115.

McInnes, L., Healy, J., and Melville, J., 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv e-prints*, arXiv:1802.03426.

McMahon, R.G., Banerji, M., Gonzalez, E., et al., 2013. First Scientific Results from the VISTA Hemisphere Survey (VHS). *The Messenger*, 154:35.

Mennickent, R.E., Garcés, J., Djurašević, G., et al., 2020. Long photometric cycle and disk evolution in the β Lyrae-type binary OGLE-BLG-ECL-157529. *A&A*, 641:A91.

Miglio, A., Montalbán, J., and Dupret, M.A., 2007. Instability strips of slowly pulsating B stars and β Cephei stars: the effect of the updated OP opacities and of the metal mixture. *MNRAS*, 375(1):L21.

Minniti, D., Contreras Ramos, R., Zoccali, M., et al., 2016. Discovery of RR Lyrae Stars in the Nuclear Bulge of the Milky Way. *ApJ*, 830(1):L14.

Minniti, D., Lucas, P., Emerson, J., et al., 2010. Vista variables in the via lactea (vvv): The public eso near-ir variability survey of the milky way. *New Astronomy*, 15(5):433.

Molnar, T.A., Sanders, J.L., Smith, L.C., et al., 2022. Variable star classification across the Galactic bulge and disc with the VISTA Variables in the Vía Láctea survey. *MNRAS*, 509(2):2566.

Montmerle, T., 1990. The Close Circumstellar Environment of Young Stellar Objects. In G. Klare, editor, *Accretion and Winds*, pages 209–233. Springer Berlin Heidelberg, Berlin, Heidelberg.

Morningstar, W.R., Levasseur, L.P., Hezaveh, Y.D., et al., 2019. Data-driven reconstruction of gravitationally lensed galaxies using recurrent inference machines. *The Astrophysical Journal*, 883(1):14.

Mowlavi, N., Holl, B., Lecoeur-Taïbi, I., et al., 2023. Gaia Data Release 3. The first Gaia catalogue of eclipsing-binary candidates. *A&A*, 674:A16.

Naul, B., Bloom, J.S., Pérez, F., et al., 2018. A recurrent neural network for classification of unevenly sampled variable stars. *Nature Astronomy*, 2(2):151.

Neumann, J.V., 1941. Distribution of the Ratio of the Mean Square Successive Difference to the Variance. *The Annals of Mathematical Statistics*, 12(4):367.

Odewahn, S.C., Stockwell, E.B., Pennington, R.L., et al., 1992. Automated Star/Galaxy Discrimination With Neural Networks. *AJ*, 103:318.

Park, W., Lee, J.E., Peña, C.C., et al., 2021. Quantifying variability of young stellar objects in the mid-infrared over 6 years with the near-earth object wide-field infrared survey explorer. *The Astrophysical Journal*, 920(2):132.

Parker, C.S., Parsons, S., Bandy, J., et al., 2019. From invisibility to readability: Recovering the ink of herculaneum. *PLOS ONE*, 14(5):1.

Paszke, A., Gross, S., Massa, F., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pawlak, M., Pejcha, O., Jakubčík, P., et al., 2019. The ASAS-SN catalogue of variable stars - IV. Periodic variables in the APOGEE survey. *MNRAS*, 487(4):5932.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al., 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825.

Pennington, R.L., Humphreys, R.M., Odewahn, S.C., et al., 1993. The Automated Plate Scanner Catalog of the Palomar Sky Survey I. Scanning Parameters and Procedures. *PASP*, 105:521.

Perera, M.V., Nair, N.G., Bandara, W.G.C., et al., 2023. SAR Despeckling Using a Denoising Diffusion Probabilistic Model. *IEEE Geoscience and Remote Sensing Letters*, 20:3270799.

Plavchan, P., Jura, M., Kirkpatrick, J.D., et al., 2008. Near-Infrared Variability in the 2MASS Calibration Fields: A Search for Planetary Transit Candidates. *ApJ*, 175(1):191.

Raddick, M.J., Bracey, G., Gay, P.L., et al., 2010. Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. *Astronomy Education Review*, 9(1):010103.

Ramesh, A., Pavlov, M., Goh, G., et al., 2021. Zero-Shot Text-to-Image Generation. *arXiv e-prints*, arXiv:2102.12092.

Ramos, R.C., Minniti, D., Gran, F., et al., 2018. The vvv survey rr lyrae population in the galactic center region*. *The Astrophysical Journal*, 863(1):79.

Rasmussen, C.E. and Williams, C.K.I., 2006. *Gaussian Processes for Machine Learning*.

Richards, J.W., Starr, D.L., Butler, N.R., et al., 2011. On machine-learned classification of variable stars with sparse and noisy time-series data. *The Astrophysical Journal*, 733(1):10.

Ricker, G.R., Winn, J.N., Vanderspek, R., et al., 2015. Transiting Exoplanet Survey Satellite (TESS). *Journal of Astronomical Telescopes, Instruments, and Systems*, 1:014003.

Riess, A.G., Casertano, S., Yuan, W., et al., 2019. Large Magellanic Cloud Cepheid Standards Provide a 1% Foundation for the Determination of the Hubble Constant and Stronger Evidence for Physics beyond ΛCDM. *ApJ*, 876(1):85.

Rimoldini, L., Holl, B., Audard, M., et al., 2019. Gaia Data Release 2. All-sky classification of high-amplitude pulsating stars. *A&A*, 625:A97.

Rimoldini, L., Holl, B., Gavras, P., et al., 2023. Gaia Data Release 3. All-sky classification of 12.4 million variable sources into 25 classes. *A&A*, 674:A14.

Robitaille, T.P., 2008. SED Modeling of Young Massive Stars. In H. Beuther, H. Linz, and T. Henning, editors, *Massive Star Formation: Observations Confront Theory*, volume 387 of *Astronomical Society of the Pacific Conference Series*, page 290.

Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386.

Saito, R.K., Hempel, M., Minniti, D., et al., 2012. VVV DR1: The first data release of the Milky Way bulge and southern plane from the near-infrared ESO public survey VISTA variables in the Vía Láctea. *A&A*, 537:A107.

Saleh, A., Sheaves, M., Jerry, D., et al., 2022. Transformer-based Self-Supervised Fish Segmentation in Underwater Videos. *arXiv e-prints*, arXiv:2206.05390.

Samus', N.N., Kazarovets, E.V., Durlevich, O.V., et al., 2017. General catalogue of variable stars: Version GCVS 5.1. *Astronomy Reports*, 61(1):80.

Sanders, J.L., Matsunaga, N., Kawata, D., et al., 2022. Mira variables in the Milky Way's nuclear stellar disc: discovery and classification. *MNRAS*, 517(1):257.

Sarmiento, R., Huertas-Company, M., Knapen, J.H., et al., 2021. Capturing the physics of manga galaxies with self-supervised machine learning. *The Astrophysical Journal*, 921(2):177.

Sarro, L.M., Debosscher, J., Aerts, C., et al., 2009. Comparative clustering analysis of variable stars in the Hipparcos, OGLE Large Magellanic Cloud, and CoRoT exoplanet databases. *A&A*, 506(1):535.

Sasaki, H., Willcocks, C.G., and Breckon, T.P., 2021. UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models. *arXiv e-prints*, arXiv:2104.05358.

Savino, A., Koch, A., Prudil, Z., et al., 2020. The age of the Milky Way inner stellar spheroid from RR Lyrae population synthesis. *A&A*, 641:A96.

Scargle, J.D., 1982. Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. *ApJ*, 263:835.

Schechter, P.L., Mateo, M., and Saha, A., 1993. DoPHOT, A CCD Photometry Program: Description and Tests. *PASP*, 105:1342.

Schwarzenberg-Czerny, A., 1996. Fast and Statistically Optimal Period Search in Uneven Sampled Observations. *ApJ*, 460:L107.

Schwarzenberg-Czerny, A., 1999. Optimum Period Search: Quantitative Analysis. *ApJ*, 516(1):315.

Sen, K., Langer, N., Pauli, D., et al., 2023. Reverse Algols and hydrogen-rich Wolf-Rayet stars from very massive binaries. *A&A*, 672:A198.

Shen, H., George, D., Huerta, E.A., et al., 2019. Denoising gravitational waves with enhanced deep recurrent denoising auto-encoders. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3237–3241.

Skarka, M., 2014. Bright Blazhko RRab Lyrae stars observed by ASAS and the SuperWASP surveys. *A&A*, 562:A90.

Skowron, D.M., Skowron, J., Mróz, P., et al., 2019. A three-dimensional map of the Milky Way using classical Cepheid variable stars. *Science*, 365(6452):478.

Slijepcevic, I.V., Scaife, A., Walmsley, M., et al., 2022. Learning useful representations for radio astronomy "in the wild" with contrastive learning. In *Machine Learning for Astrophysics*, page 53.

Smith, L.C., Lucas, P.W., Kurtev, R., et al., 2018. VIRAC: the VVV Infrared Astrometric Catalogue. *MNRAS*, 474(2):1826.

Smith, M.J., Fleming, L., and Geach, J.E., 2023. EarthPT: a time series foundation model for Earth Observation. *arXiv e-prints*, arXiv:2309.07207.

Smith, M.J. and Geach, J.E., 2019. Generative deep fields: arbitrarily sized, random synthetic astronomical images through deep learning. *MNRAS*, 490(4):4985.

Sohn, K., 2016. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29 of *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Sokolovsky, K.V., Gavras, P., Karampelas, A., et al., 2017. Comparative performance of selected variability detection techniques in photometric time series data. *MNRAS*, 464(1):274.

Soszyński, I., Pawlak, M., Pietrukowicz, P., et al., 2016. The OGLE Collection of Variable Stars. Over 450 000 Eclipsing and Ellipsoidal Binary Systems Toward the Galactic Bulge. , 66(4):405.

Soszyński, I., Udalski, A., Wrona, M., et al., 2019. Over 78 000 RR Lyrae Stars in the Galactic Bulge and Disk from the OGLE Survey. , 69(4):321.

Srivastava, N., Hinton, G., Krizhevsky, A., et al., 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929.

Stassun, K.G. and Torres, G., 2021. Parallax systematics and photocenter motions of benchmark eclipsing binaries in gaia edr3. *The Astrophysical Journal Letters*, 907(2):L33.

Stellingwerf, R.F., 1978. Period determination using phase dispersion minimization. *ApJ*, 224:953.

Stetson, P.B., 1996. On the Automatic Determination of Light-Curve Parameters for Cepheid Variables. *PASP*, 108:851.

Storrie-Lombardi, M.C., Lahav, O., Sodré, L., J., et al., 1992. Morphological Classification of galaxies by Artificial Neural Networks. *Monthly Notices of the Royal Astronomical Society*, 259(1):8P.

Sulc, A., Kammering, R., Eichler, A., et al., 2023. PACuna: Automated Fine-Tuning of Language Models for Particle Accelerators. *arXiv e-prints*, arXiv:2310.19106.

Sun Jin, Wu Shi-min, and Fan Ying, 1982. A further discussion on the mass loss of mira stars. *Chinese Astronomy and Astrophysics*, 6(2):104.

Swingler, D.N., 1989. A comparison of the Fourier, Jurkevich, and Stellingwerf methods of period estimation. *AJ*, 97:280.

Takeuti, M., Nakagawa, A., Kurayama, T., et al., 2013. A Method to Estimate the Masses of Asymptotic Giant Branch Variable Stars. , 65:60.

Taylor, M.B., 2005. TOPCAT & STIL: Starlink Table/VOTable Processing Software. In P. Shopbell, M. Britton, and R. Ebert, editors, *Astronomical Data Analysis Software and Systems XIV*, volume 347 of *Astronomical Society of the Pacific Conference Series*, page 29.

Templeton, M.R., Mattei, J.A., and Willson, L.A., 2005. Secular Evolution in Mira Variable Pulsations. *AJ*, 130(2):776.

Udalski, A., Szymański, M.K., and Szymański, G., 2015. OGLE-IV: Fourth Phase of the Optical Gravitational Lensing Experiment. , 65(1):1.

Vagnetti, F., Middei, R., Antonucci, M., et al., 2016. Ensemble X-ray variability of active galactic nuclei. II. Excess variance and updated structure function. *A&A*, 593:A55.

van der Maaten, L. and Hinton, G., 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579.

van Terwisga, S.E., Hacar, A., van Dishoeck, E.F., et al., 2022. Survey of Orion Disks with ALMA (SODA). I. Cloud-level demographics of 873 protoplanetary disks. *A&A*, 661:A53.

VanderPlas, J.T., 2018. Understanding the Lomb-Scargle Periodogram. *ApJ*, 236(1):16.

Vaswani, A., Shazeer, N., Parmar, N., et al., 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010. Curran Associates Inc., Red Hook, NY, USA.

Vaughan, S., Edelson, R., Warwick, R.S., et al., 2003. On characterizing the variability properties of X-ray light curves from active galaxies. *MNRAS*, 345(4):1271.

Vio, R., Andreani, P., and Biggs, A., 2010. Unevenly-sampled signals: a general formalism for the Lomb-Scargle periodogram. *A&A*, 519:A85.

von Hippel, T., Storrie-Lombardi, L.J., Storrie-Lombardi, M.C., et al., 1994. Automated Classification of Stellar Spectra - Part One - Initial Results with Artificial Neural Networks. *MNRAS*, 269:97.

Walmsley, M., Smith, L., Lintott, C., et al., 2020. Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning. *MNRAS*, 491(2):1554.

Wang, Y., Khardon, R., and Protopapas, P., 2012. Nonparametric Bayesian Estimation of Periodic Light Curves. *ApJ*, 756(1):67.

Wehenkel, A., Behrmann, J., Miller, A.C., et al., 2023. Simulation-based Inference for Cardiovascular Models. *arXiv e-prints*, arXiv:2307.13918.

Wills, S., Underwood, C.J., and Barrett, P.M., 2023. Machine learning confirms new records of maniraptoran theropods in middle jurassic uk microvertebrate faunas. *Papers in Palaeontology*, 9(2):e1487. E1487 PALA-05-22-5317-OA.R1.

Wolk, S.J., Günther, H.M., Poppenhaeger, K., et al., 2018. YSOVAR: Mid-infrared Variability among YSOs in the Star Formation Region Serpens South. *AJ*, 155(2):99.

Wolk, S.J., Rice, T.S., and Aspin, C., 2013. Near-infrared Variability among Young Stellar Objects in the Star Formation Region Cygnus OB7. *ApJ*, 773(2):145.

Wood, P.R., 1979. Pulsation and mass loss in Mira variables. *ApJ*, 227:220.

Wood, P.R. and Bessell, M.S., 1983. Long-period variables in the galactic bulge : evidence for a young super-metal-rich population. *ApJ*, 265:748.

Wright, E.L., Eisenhardt, P.R.M., Mainzer, A.K., et al., 2010. The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance. *AJ*, 140(6):1868.

Yang, R., Srivastava, P., and Mandt, S., 2023. Diffusion Probabilistic Modeling for Video Generation. *Entropy*, 25(10):1469.

York, D.G., Adelman, J., Anderson, J. E., J., et al., 2000. The Sloan Digital Sky Survey: Technical Summary. *AJ*, 120(3):1579.

Zechmeister, M. and Kürster, M., 2009. The generalised Lomb-Scargle periodogram. A new formalism for the floating-mean and Keplerian periodograms. *A&A*, 496(2):577.

Zhang, R. and Zou, Q., 2018. Time Series Prediction and Anomaly Detection of Light Curve Using LSTM Neural Network. In *Journal of Physics Conference Series*, volume 1061 of *Journal of Physics Conference Series*, page 012012.

Zhu, W.W., Berndsen, A., Madsen, E.C., et al., 2014. Searching for pulsars using image pattern recognition. *The Astrophysical Journal*, 781(2):117.