Extreme Binaries in a Target Rich Environment

Dharmesh Mistry

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

February 2025

Declaration

The work presented in this thesis was carried out at the Astrophysics Research Institute, Liverpool John Moores University. Unless otherwise stated, it is the original work of the author.

While registered as a candidate for the degree of Doctor of Philosophy, for which submission is now made, the author has not been registered as a candidate for any other award. This thesis has not been submitted in whole, or in part, for any other degree.

DHARMESH MISTRY
Astrophysics Research Institute
Liverpool John Moores University
146 Brownlow Hill
Liverpool
L3 5RF
UK

February 2025

Abstract

Cataclysmic Variables (CVs) are a class of binary star systems consisting of a white dwarf accreting matter from a main sequence companion star. The diversity in their physical properties causes them to exhibit a wide range of observable phenomena, including dwarf nova outbursts and novae explosions, making them crucial for the study of binary evolution and accretion physics. The discovery and characterisation of CVs have been greatly facilitated by wide-field time domain surveys, such as the Catalina Real-Time Transient Survey (CRTS) and the Zwicky Transient Facility (ZTF). They can detect significant changes in the brightness of astrophysical objects on various timescales to generate alerts. Due to the continuing advancements in survey technology, alert rates are on the rise, with the Rubin Observatory expected to generate of order 10^7 alerts per night. Given the large alert rates, classifying these events by the class of astrophysical transient responsible can no longer be solely performed manually. Moreover, follow-up facilities are too few to characterise all events, therefore, follow-up time will be reserved for the rarest of events.

Source classification and the search for the rarest of events in this deluge of alerts requires the automation provided by Machine Learning (ML). ML-based source classification is an active research field, with the distinction between many classes of transient possible (e.g., supernovae, active galactic nuclei, and variable star subtypes). However, ML-based searches for CVs is an underdeveloped field; furthermore, an emphasis on the identification/classification of CV subtypes with ML is unexplored. Given the diversity present within the CV transient class and the under-representation of certain subtypes, such as those with strongly magnetic white dwarfs and ultra-short period helium accreting CVs, a ML-based pipeline specifically purposed for such a task is much needed. The objective of this research has been to address this gap, whilst also identifying the factors that hinder this objective.

On this pathway transient sources published by the Gaia Science Alerts program (GSA) were explored with ML. Utilising the technique of light curve feature extraction and with the aid of source metadata from the Gaia survey a ML model based on the Random Forest algorithm was produced. It is capable of distinguishing CVs from supernovae, active galactic nuclei and young stellar objects with a 92% precision score (fraction of those predicted as CV belonging to the class). Of 13,280 sources within GSA without an assigned transient classification, the model predicts the CV class for ∼2800, of which spectroscopic confirmation has been acquired for 15 so far.

During the next research phase, the higher cadence, multi-band survey of the Zwicky Transient Facility (ZTF) was explored. A two-stage ML pipeline was developed that comprises and alerts filtering stage aimed at removing non-CVs, followed by an ML classifier tasked with dividing the filtered sources into their CV subtypes based on features extracted from their light curves in combination with Gaia DR3 data. During the month of June 2023 alone, 51 candidates of the CV class were discovered, 14 of which are candidates of either the rare AM CVn or polar CV subtypes. Representations of the ML classifier's prediction patterns, input into the Generative Topographic Mapping algorithm, indicate the influence of CV evolutionary factors. CV evolution and the consequential blending of boundaries that separate CV subtypes from one another, is found to be a major factor in the difficulty of distinguishing between CV subtypes.

To conclude the research, dimensionality reduction techniques were explored with the ZTF dataset. The findings reaffirm the view that CV evolution plays a major factor in the difficulty in distinguishing between subtypes, as do the intricacies of the ZTF survey photometry. In addition, the reduced dimensionality representations were found to be particularly valuable in approximating a subtype classification, with distinct locations of strongly eclipsing CVs as well as polars a particular highlight.

Acknowledgements

I would like to express my deepest gratitude to my lead supervisor, Chris Copperwheat. Your knowledge, guidance, and support have been invaluable throughout my research journey. I am also grateful to my co-supervisors, Ivan Olier and Matt Darnley. Ivan, your expertise in machine learning has been indispensable, as has your sense of humour (and thank you for the postdoc opportunity, by the way). Matt, your keen insight and attention to detail have been invaluable assets — you're now cited for '12a'; I can't believe I missed that!

I would also like to thank Prof. Iain Steele, who guided me through my Master's project and encouraged me to pursue a PhD. Thanks as well to my examiners, Dan Perley and Simone Scaringi, whose suggestions have helped improve this thesis.

I would like to thank the Faculty of Engineering and Technology for their financial support, which enabled me to pursue my research endeavours. I am grateful to everyone from the Astrophysics Research Institute who has helped along the way. Special thanks go to all members of room 2.30, former and current, who had to put up with me during my academic journey, the supportive environment you created has undoubtedly impacted my academic journey in ways I may not fully comprehend.

I would like to extend my heartfelt thanks to my parents, Tara and Magan. Although they passed away before my academic journey, I believe they must have instilled values in me that have contributed to my successes thus far. Last but not least, I'd also like to thank my brothers Umesh and Kailesh, as well as my close relatives for their unwavering support.

DHARMESH MISTRY February 2025

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cataclysmic variables (CVs; Warner 1995; Hellier 2001) are compact accreting binary systems consisting of a white dwarf accreting matter from a Roche-lobe filling donor. In most cases, the donor is a low-mass, late-type, main-sequence star. CVs are a large and easily observed population, providing ideal laboratories for the study of binary evolution and the physics of accretion. For example, they represent a possible single degenerate pathway towards type Ia supernovae (SNe Ia) — events that serve as 'standard candles' to determine distances to other galaxies and constrain our cosmological models (Phillips, 1993). A greater understanding of CV evolution may elucidate our understanding of the multiple progenitor scenarios believed to produce those SNe Ia events that deviate from the 'normal' (Jha et al., 2019). As laboratories for the study of accretion, CVs probe extreme conditions. For example, the white dwarf in $\sim 20-25\%$ of CVs (Ferrario et al., 2015) possesses a magnetic field sufficiently strong to divert the flow of matter arriving from the donor star out of the orbital plane before accretion onto the white dwarf's magnetic poles (Cropper, 1990; Patterson, 1994). The AM CVn subclass of CVs (Solheim, 2010) are ideal for examining helium accretion, whilst also serving as laboratories for examining accretion from a semi-degenerate/degenerate companion, and accretion occurring at ultrashort periods (between 5 and 65 minutes).

One of the main methods of CV discovery is via their photometric variation. Recent decades have seen a dramatic increase in the potential for their discovery due to the development of wide field, high cadence and panchromatic surveys that repeatedly image huge areas of the sky at high rates. Surveys include the (intermediate) Palomar Transient

Factory (iPTF; Cao et al. 2016), the Optical Gravitational Lensing Experiment (OGLE; Udalski et al. 2015), and newer surveys such as Catalina Real-time Transient Survey (CRTS; Drake et al. 2009), the Zwicky Transient Facility (ZTF; Bellm et al. 2019), the All-Sky Automated Survey for Supernovae (ASAS-SN; Shappee et al. 2014), and Gaia Science Alerts (GSA; Hodgkin et al. 2021). They adopt the difference imaging technique (Kerins et al., 2010), in which the new sky image is subtracted from a reference image, to reveal objects that have changed in brightness. Plots of brightness over time (light curves) of said objects may be used to deduce the astrophysical transient class responsible. Examples of the effectiveness of these surveys include the detection of 705 new CV candidates from 5 years of CRTS data (Drake et al., 2014), the 497 CV candidates uncovered from two years of ZTF transients (Szkody et al., 2020, 2021), and nine outbursting examples of the AM CVn CV subclass found by van Roestel et al. (2021), also from ZTF transients.

These examples involve a heavy focus on the manual inspection of large amounts of data to distinguish CVs from other time-varying sources (e.g., supernovae, variable stars, and active galactic nuclei). This practice is becoming ever more time-consuming and infeasible due to the large numbers of transient events reported every night, events that also include artefacts from difference imaging (e.g., poorly subtracted galaxies, cosmic rays, and defective pixels (Goldstein et al., 2015). In the case of ZTF, transient alert rates can exceed a million per night (Patterson et al., 2019), a rate set to be dwarfed by the Rubin Observatory (Ivezić et al., 2019). As a further side effect, facilities devoted to the follow-up of transient events are not enough in number to investigate them all, therefore time on such facilities is limited. Since the majority of genuine astrophysical sources may serve only to reaffirm our current understanding of the transient classes to which they belong, follow-up time will be reserved for the minority, those that present a challenge to or help further our understanding.

Machine learning (ML) is ideally suited to address these challenges, with the vetting of artefacts now heavily reliant on automated pipelines (e.g., Goldstein et al. 2015; van Roestel et al. 2021). Applications for transient source identification/classification are becoming ever more abundant. For example, van Roestel et al. (2021) describes the ZTF Source Classification Project, a hierarchical ML pipeline that aims to group ZTF alerting transients into both variability types and transient classes that include Active Galactic Nuclei (AGN), young stellar objects (YSOs), variable stars and CVs based on

photometry. Neira et al. (2020) generated ML classifiers to distinguish between AGN, Blazars, CVs, supernovae, and non-transients (amongst others) from CRTS source light curves. Alert brokers have been used by ZTF to ingest and classify alerts, serving them to the astronomical community. The ALeRCE (Automatic Learning for the Rapid Classification of Events; Förster et al. 2021) broker makes use of science, reference and difference images for rapid classification of events (Carrasco-Davis et al., 2021), and multiband light curves for classification of events with longer-term observations into several transient classes including CVs, supernova subtypes, AGN, and variable stars. Sun et al. (2021) focused on spectroscopic data, searching for CVs within Data Release 6 (DR6) of the LAMOST survey (Cui et al., 2012) containing nearly 10 million low-resolution spectra.

These examples focus on the identification/classification of CVs as a broad class. However, CVs are a diverse class of accreting binaries, with a correspondingly diverse range of photometric (and spectroscopic) behaviour, therefore significant human vetting is still required to make these important distinctions. The ability to automatically group CVs into their respective subtypes, and/or identify rare subtypes remains an underdeveloped research field. Such a classifier/pipeline should dramatically reduce the time required for human vetting, which will become all the more important as transient source detection capabilities improve with time. The following work describes the journey towards the creation of such a pipeline that aims to serve the CV research community with a regular supply of interesting candidates. However, before taking you on this journey, I use Chapter 2 to provide an extensive overview of CVs, including their diverse subtypes and significance in astrophysical research. Chapter 3 delves into the machine learning techniques and methodologies used in this study, while in Chapter 4, I discuss the importance of transient surveys in CV research, the impact of machine learning in managing survey data, and the gap in source classification research that I aim to fill. These chapters set the stage for the subsequent focus on developing an automated source classification pipeline tailored to the CV community's needs.

Chapter 2

Cataclysmic variables

2.1 Introduction

Cataclysmic variables (Warner, 1995; Hellier, 2001) are a class of binary star systems consisting of a white dwarf (WD or primary) and typically a low-mass main sequence star (donor or secondary). The binary components orbit very close to one another with separations of less than a few solar radii and orbital periods typically less than half a day. The strong gravitational pull of the WD causes the donor to transfer matter to the primary via Roche lobe overflow at rates ranging from 10^{-11}–10^{-8} $M_{\odot}$ yr^{-1} (Patterson, 1984). The hydrogen-rich matter from the donor star is accreted onto the WD surface via, in most cases, an accretion disk which forms around the WD. Alternatively, if the magnetic field of the WD is sufficiently strong, the formation of an accretion disk is either completely or partially inhibited, with the matter flowing along the field lines to the magnetic poles of the WD — which occurs for $\sim 20-25\%$ of the observed CV population (Ferrario et al., 2015). Focusing on the non-magnetic case for simplicity (Figure 2.1), the point of interaction between the incoming stream of matter from the donor and the outer accretion disk is referred to as the bright spot. The boundary layer is defined as the region between the inner accretion disk and the WD surface. The donor, the WD, accretion disk, bright spot and boundary layer each contribute to emission that spans the electromagnetic spectrum from X-ray to infrared.

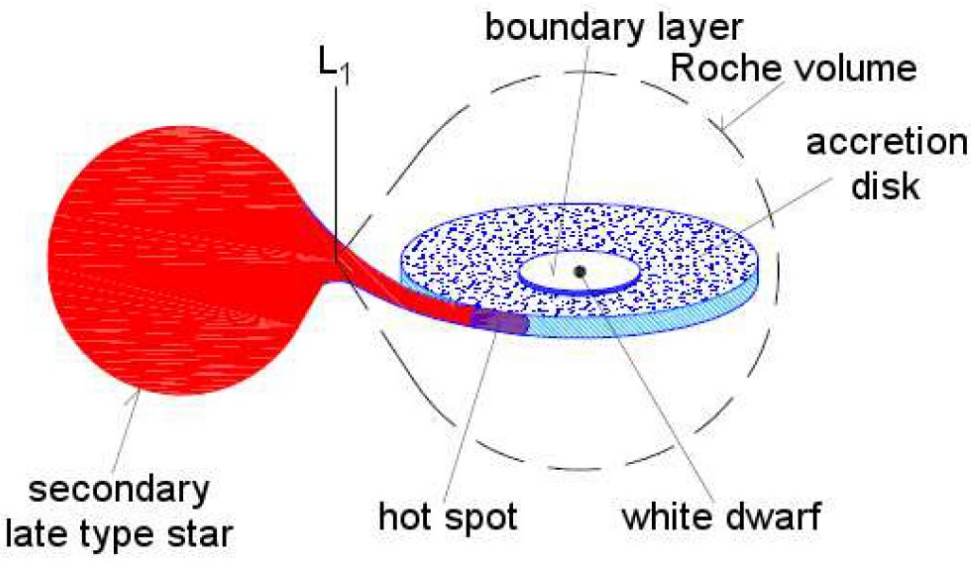


FIGURE 2.1: Taken from Giovannelli (2008), this is a sketch of a non-magnetic CV with all components responsible for the energy emission: WD, donor star, accretion disk, bright spot, and boundary layer.

2.2 Emission Components

White dwarf

White dwarfs are the end point of stellar evolution for stars of main-sequence mass of $10\,M_\odot$ or less. They are the hot dense cores of their main sequence progenitors, supported against their gravity by electron degeneracy pressure. Pala et al. (2021) found that CV white dwarf masses are within the range $0.35\,M_\odot$ and $1.25\,M_\odot$, with a mean average of $0.81\,M_\odot$, greater than that of single white dwarfs ($0.6\,M_\odot$). CV white dwarfs tend to be hotter at longer periods, reflecting the higher rates of accretion. Their effective surface temperatures generally lie within the range $\sim 12000\,K$ and $\sim 40000\,K$ with radii approximately that of the Earth. The peak of their spectral energy distribution occurs at far-ultraviolet wavelengths.

Donor star

The donor stars of CVs are typically red dwarf stars with spectral types of late K or M type. The spectral type is strongly dependent on the orbital period, where shorter period systems are typically of a later spectral type (Knigge, 2006; Knigge et al., 2011). This is shown in Figure 2.2, where it can also be seen that the spectral types of CV

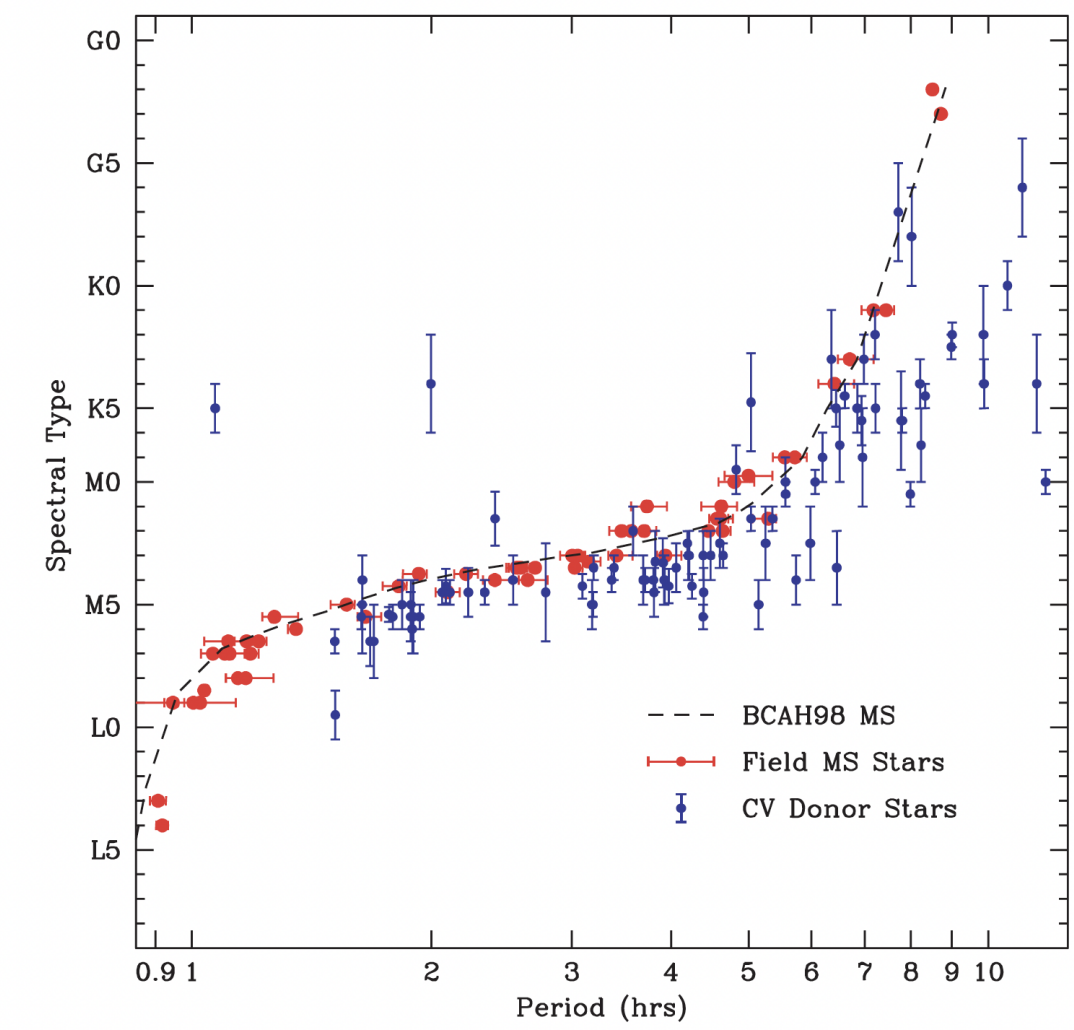


FIGURE 2.2: Spectral Type versus Orbital Period for CV donors and single main sequence stars. The MS stars are assigned an orbital period based on their mass and radius via the period-density relation for Roche lobe filling stars (Equation 2.10). The blue dots represent CV donors whose spectral types are empirically derived, while the red dots are the main sequence stars. Taken from Knigge (2006)

donors are systematically later than isolated main sequence stars of the same mass. This is believed to be due to mass loss driving donors slightly out of thermal equilibrium, causing them to be bloated compared to MS stars of the same mass. Donor radii span a range from $\sim 0.1\,R_\odot$ to $\sim 0.6\,R_\odot$, while masses are in the range $\sim 0.05\,M_\odot$ to $\sim 0.6\,M_\odot$ (excluding evolved donors). CV donors are cool, with effective temperatures typically below 4,000 K, although they can range from 500 K to 4200 K (Knigge et al., 2011). One side of the donor is heated by the hot WD causing the temperature on this side to reach $\sim$ 7,500 K, which may result in smooth brightness variations in high inclination systems.

Accretion Disks

Accretion disks (Frank et al., 2002) are usually responsible for the majority of optical emission. The surface temperature, $T_{surface}$, of accretion disks is a function of the disk radius r, and approximately follows $T_{surface} \propto r^{-3/4}$. One may model the emission of the accretion disk at various frequencies by assuming each disk annulus emits as a blackbody of a particular temperature and summing the blackbody contributions from each annulus (weighted by its area).

Optical lines (usually in emission) in CV spectra are believed to be of disk origin with broad Balmer emission usually most prominent. Under optically thick disk conditions, absorption may be observed, while optically thin conditions give rise to broad Balmer emission. Double-peaked emission profiles may be observed with the extent of the feature more pronounced at higher orbital inclinations. This can be understood by considering the disk as a collection of small emission regions. As matter in the disk circles the WD, the emission from one half of the disk will be blue-shifted, whilst that from the other half will be red-shifted. The combination of decreasing velocity with increasing radius and the smaller disk area at smaller radii gives rise to the profile shown and further explained in Figure 2.3.

Bright Spot

This is the region where the matter stream from the donor impacts the outer rim of the accretion disk (Hellier, 2001; Warner, 1995). The impact, at supersonic speeds, shock-heats the region and radiates possibly as much or more energy in the optical than the other components combined (primary, donor, disk) via thermal blackbody emission. This can be deduced through observations of orbital humps present in light curves of high inclination systems, for example, IY UMa in Figure 2.4. The humps are caused by our changing view of the bright spot throughout the orbit (Patterson et al., 2000).

Boundary Layer

The Keplerian velocity just above the white dwarf surface ($\sim$ 3000 km/s) is much faster than the white dwarf surface rotation speed ($\sim$ 300 km/s). Therefore, matter in the inner disk decelerates to match the white dwarf rotation. This transition region is known as the boundary layer (Frank et al., 2002). The kinetic energy of the slowing matter is converted to heat and radiated away through Bremsstrahlung radiation (Mukai, 2017).

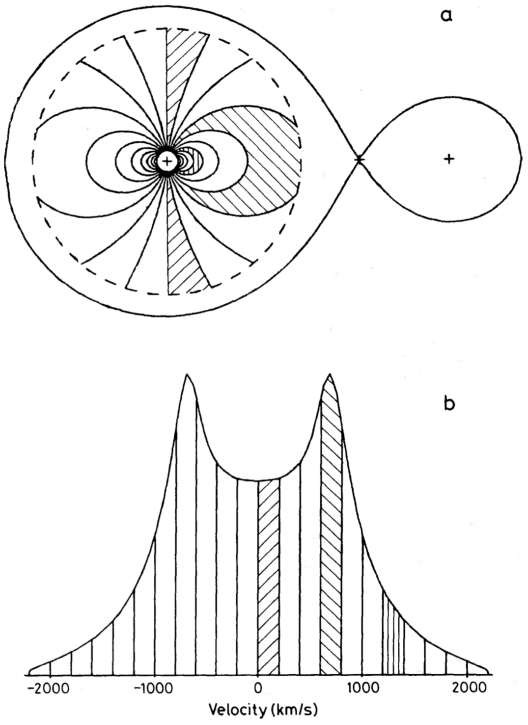


FIGURE 2.3: (a) A Keplerian accretion disk with different velocity regions as viewed by an observer situated below the plot. (b) The resultant double-peaked profile. Emission in the shaded velocity bins arise from the corresponding regions of the disk in (a). The highest velocity regions in the disk produce the lowest emission due to their low surface area. The lowest velocity regions originate from disk material aligned with our line of sight, moving tangentially to it.

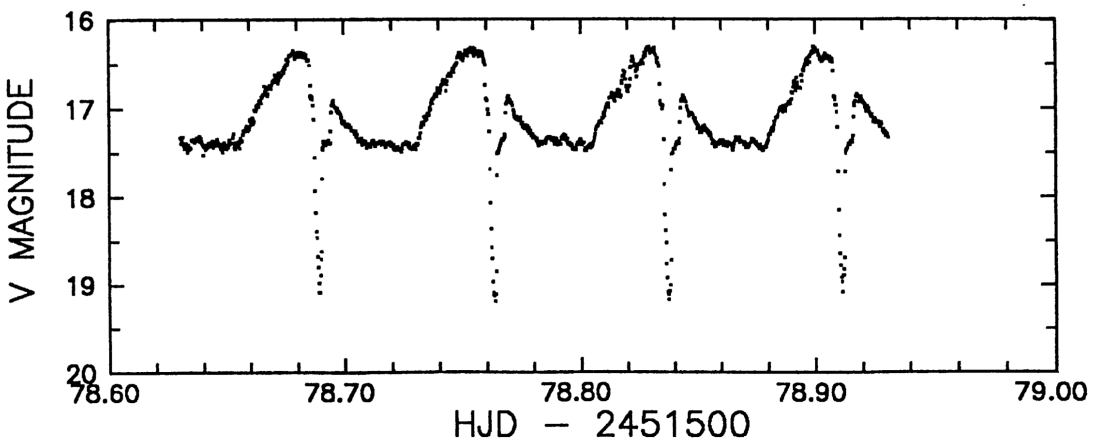


FIGURE 2.4: A light curve of IY UMa showing 'orbital humps' due to the changing view of the bright spot. Also present are the deep eclipses, with orbital hump peaking before each eclipse (Patterson et al., 2000).

At high accretion rates, the layer becomes optically thick, emitting as a blackbody at temperatures of $\sim 200,000$ K (extreme UV to soft X-ray). The situation alters at low accretion rates ($\leq\sim 5 \times 10^{10}\,\mathrm{M_{\odot}yr^{-1}}$); the high-temperature matter is too diffuse to efficiently radiate energy and cool down. In response, the matter expands, making it even less capable of cooling, causing further expansion and the evaporation of the inner disk into a hot ($\sim 10^8$ K), optically thin, diffuse 'corona' that emits hard X-rays. The transition between high and low accretion rates occurs during dwarf nova outbursts (see Section 2.4.1). During an outburst a high accretion rate is present, the boundary layer is optically thick and emits extreme UV/soft X-rays. In quiescence (low accretion rate), the boundary layer evaporates into a corona producing hard X-ray emission. The anti-correlation between extreme UV/soft X-ray and hard X-ray emission during outbursts has been observed in SS Cyg (Wheatley et al., 2003). The optically thin corona that flows outwards over the disk is believed to be a major source of emission lines in dwarf nova spectra during quiescence.

Spectral energy distribution

The spectral energy distribution of CVs comprises contributions from each of the aforementioned emission components. At high energies, the emission will originate from the boundary layer at X-ray to extreme UV wavelengths, while the white dwarf contribution will be largely concentrated in the ultraviolet with a diminishing contribution towards longer wavelengths. However, if the boundary layer is optically thick, white dwarf emission is obscured by boundary layer emission. The red dwarf emits most strongly in the infrared. In the majority of cases, the accretion disk will dominate the spectrum in the optical, with higher and lower energy emission generated from the inner and outer disk edges respectively. The bright spot contribution will depend on the brightness, temperature and orbital inclination, though emission will largely be confined to the optical. The combined emission from these components from the infrared to UV will usually resemble a blackbody flattened by the optical emission of the accretion disk. Should the contribution of one or both stars be significant in comparison to the disk, a noticeable increase in emission will appear towards both/either the UV and/or the IR regime. Figure 2.5 shows a schematic spectrum of a CV showing the contributions of the white dwarf, donor and accretion disk.

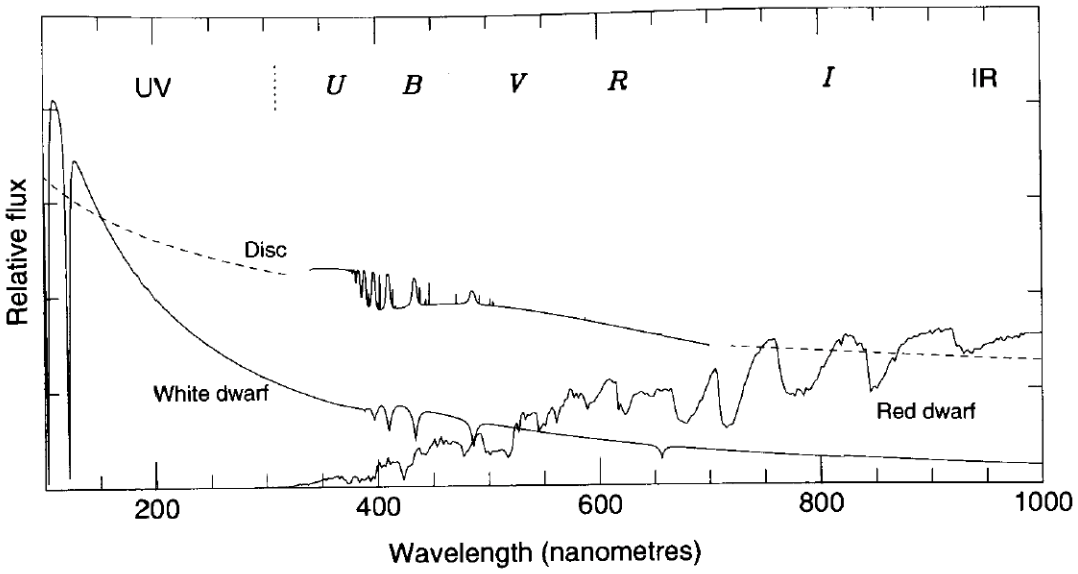


FIGURE 2.5: Schematic spectrum of a CV showing contributions from the white dwarf, red dwarf (donor) and accretion disk (Hellier, 2001).

2.3 Formation and evolution

2.3.1 Roche lobe Geometry

To aid in explaining CV formation and evolution, a description of Roche lobe geometry is required. Carroll & Ostlie (1996) and Frank et al. (2002) provide a detailed description that is summarised here. Most binary systems have orbital separations large enough such that the only interaction between the components is through their mutual gravitational attraction. However in close binaries such as cataclysmic variables, tidal forces become non-negligible, distorting the geometry of one or both stars and the transfer of matter from one star to the other. This can be understood within the context of the Roche lobe, defined as the region around a star in a binary within which orbiting material is gravitationally bound to that star.

Consider a frame of reference corotating with two stars of mass M_1 and M_2. The origin is coincident with the system's centre of mass, where a circular orbit about this origin is assumed for each star. The stars are at rest in this non-inertial frame at positions $\mathbf{r}_1$ and $\mathbf{r}_2$, with their mutual gravitational attraction balanced by the outwardly directed centrifugal forces. A test mass located at $\mathbf{r}$ will experience a potential ϕ (potential energy per unit mass) that is the sum of the gravitational potentials of each star and the

centrifugal potential due to a rotating frame of reference. The total potential experienced by the test mass at any point can be expressed in vector form as:

$$\phi(\mathbf{r}) = -\frac{GM_1}{|\mathbf{r}-\mathbf{r_1}|} - \frac{GM_2}{|\mathbf{r}-\mathbf{r_2}|} - \frac{1}{2}(\Omega \times \mathbf{r})^2 \tag{2.1}$$

The first two terms on the right refer to the gravitational potentials of each star. The third term accounts for the centrifugal potential, where the angular velocity of the binary, Ω, can be expressed using Kepler's third law as:

$$\Omega = \frac{2\pi}{P_{orb}} = \left[\frac{G(M_1+M_2)}{a^3}\right]^{1/2} \mathbf{e} \tag{2.2}$$

where P_{orb}, a, and $\mathbf{e}$ are the orbital period, binary separation, and a unit vector perpendicular to the plane of the binary, respectively. If we were to consider a test mass located on the plane of the orbit, as shown in Figure 2.6, Equation 2.1 can be expressed as:

$$\phi = -\frac{GM_1}{s_1} - \frac{GM_2}{s_2} - \frac{1}{2}\Omega^2 r^2 \tag{2.3}$$

From the law of cosines the distances s_1 and s_2 are given by:

$$s_1^2 = r_1^2 + r^2 + 2r_1 r\cos\theta \tag{2.4}$$

$$s_2^2 = r_2^2 + r^2 + 2r_2 r\cos\theta \tag{2.5}$$

Equations 2.2–2.5 in combination with $r_1 + r_2 = a$ and $M_1 r_1 = M_2 r_2$ can be used to define the gravitational potential at every point on the orbital plane. Points in space that share the same ϕ form an equipotential surface. Several such surfaces are shown in Figure 2.7 for a specific set of values for M_1, M_2, and a. The shape of the equipotential surfaces is governed entirely by the mass ratio $q \equiv M_2/M_1$, while the binary separation a accounts for the overall scale.

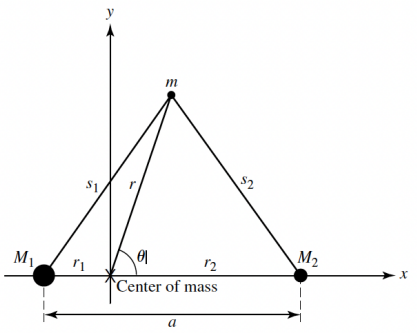


FIGURE 2.6: Taken from Carroll & Ostlie (1996). Corotating coordinates for a binary star system. The masses M_1 and M_2 are separated by a distance a. The stars are located on the x-axis at distances r_1 and r_2, respectively, from the centre of mass, which is placed at the origin.

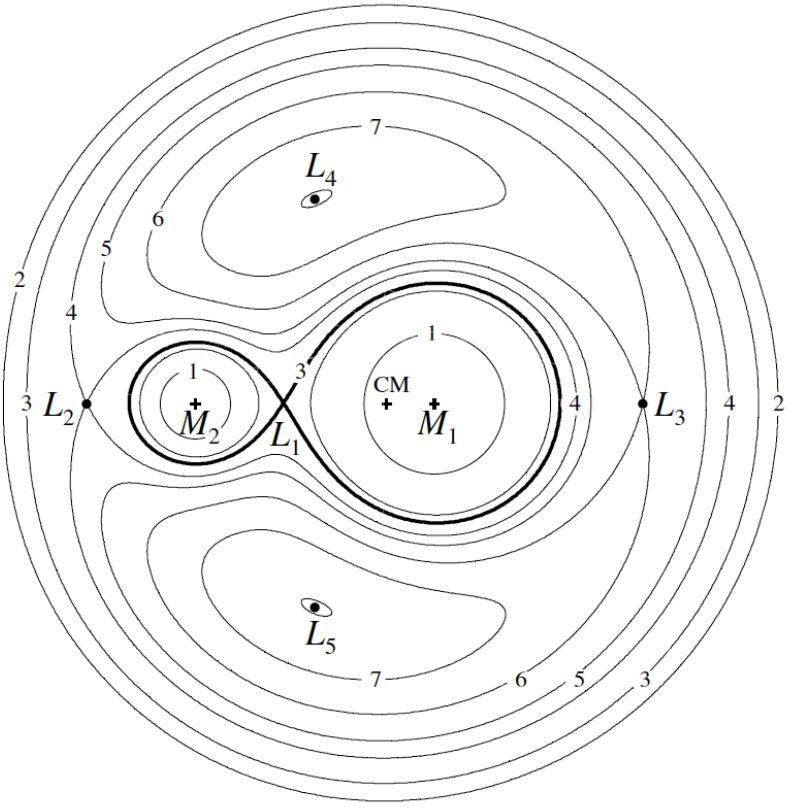


FIGURE 2.7: Taken from figure 4.3 of Frank et al. (2002). Sections in the orbital plane of the Roche equipotentials $\phi_R = constant$, for a binary system with mass ratio $q = M_2/M_1 = 0.25$. Shown are the centre of mass (CM) and Lagrange points L1–L5. The equipotential surfaces are labelled 1–7 in order of increasing ϕ_R.

The Lagrange points, L1–L5 are locations where a test mass m experiences no force, i.e., the gravitational forces on m due to M_1 and M_2 are precisely balanced by the centrifugal force. More succinctly they are balance points where $d\phi/dx = 0$. The motion of m will be dominated by the gravitational pull of the dominant star, the potential ϕ can be described as having two deep valleys at $\mathbf{r}_1$ and $\mathbf{r}_2$. As we move further from each star their shapes become distorted into 'teardrop' shapes due to the combined gravitational effects of M_1 and M_2 until they finally meet at the inner Lagrange point, L1. The surfaces that meet at L1, are the Roche lobes of each of the components, depicted by the emboldened equipotential surface forming a figure of eight in Figure 2.7. L1 can be seen as a saddle point, akin to the lowest mountain pass between the two deep valleys, it is the easiest path by which matter can pass between the two stars. At even greater distances the surfaces assume a 'dumbbell' shape, surrounding both the masses.

The appearance of a binary depends upon which equipotential surface is filled by each star. As a star evolves to larger radii, it will take on the shape of successively larger equipotential surfaces. Binary stars with radii much less than their separation take on nearly spherical shapes. These stars evolve nearly independently of one another and the binary is described as **detached**. Should one of the stars expand beyond the equipotential surface defining its Roche lobe, then material from that star may enter the lobe of its companion via the L1 Lagrangian point. Such a system is described as a **semi-detached** binary. A **contact binary** is formed when both stars fill or expand beyond their respective lobes. These stars now share a common envelope and are bounded by a dumbbell-shaped equipotential surface that encompasses both stars, e.g., that which passes through L2. Cataclysmic variables spend the majority of their lives in the semi-detached state with the less massive donor star just overfilling its Roche lobe. A consequence of this is the donor will become tidally locked to the orbital period, i.e., the donor's spin period equals the system's orbital period.

The Roche lobe treatment of binary interaction allows for a quantitative description of important binary properties, such as the correct interpretation of the shape of eclipses or the rate of mass transfer. To do this one requires a calculation of lobe geometry (e.g., sizes of the lobes and distances to the L1 point). The form of Equation 2.1 is complicated requiring numerical solutions to achieve this. However, analytical approximations are available, a few are provided as follows. Roche lobes are non-spherical, so an average radius must be found: this is done by approximating this to the radius of a sphere that

has the same volume as that of the lobe (or volume radius). The volume radius of the donor lobe, Equation 2.6, provided by Eggleton (1983) is accurate to better than 1%. It is dependent on the mass ratio, q, between the two stars, defined as $q \equiv M_2/M_1$, where M_1 and M_2 are the primary (more massive star) and secondary star masses respectively (substituting q with q^{-1} in equation 2.6 yields the volume radius for the lobe of the primary).

$$\frac{R_2}{a} = \frac{0.49q^{2/3}}{0.6q^{2/3} + ln(1 + q^{1/3})} \tag{2.6}$$

Paczyński (1971) provides a simpler and easier to interpret form of Equation 2.6 for $0.1 \leq q \leq 0.8$ (Equation 2.7).

$$\frac{R_2}{a} = \frac{2}{3^{4/3}} \left(\frac{q}{1+q} \right)^{1/3} = 0.462 \left(\frac{M_2}{M_1 + M_2} \right)^{1/3} \tag{2.7}$$

The ratio of the primary and secondary lobe volume radii is accurate to better than 5%. For mass ratios $0.03 \leq q \leq 1$ this is:

$$\frac{R_1}{R_2} = \left(\frac{M_1}{M_2} \right)^{0.45} \tag{2.8}$$

The distance b_1 of the L1 point from the centre of the primary is found using:

$$\frac{b1}{a} = 0.500 - 0.227 \log q \tag{2.9}$$

The mean density $\bar{\rho}$ of the secondary can be found solely from the orbital period P by using Equation 2.6 and Kepler's third law. For $q \leq 0.8$ this is given by:

$$\bar{\rho} = \frac{3M_2}{4\pi R_2^3} \cong \frac{3^5\pi}{8GP^2} \cong 110 P_{hr}^{-2} \,\mathrm{g\,cm^{-3}} \tag{2.10}$$

Equation 2.10 shows that for periods in the range $\sim 1 - 10$ hours, late type main sequence secondaries ($\bar{\rho} \sim 1 - 100\,\mathrm{g\,cm^{-3}}$) can fill their Roche lobes. Rearranging of Kepler's third law gives an approximation of the binary separation in terms of the mass ratio and orbital period (Equation 2.11).

$$a = 3.53 \times 10^{10} (M_1/M_\odot)^{1/3} (1+q)^{1/3} P_{orb}^{2/3}(h) \text{ cm} \tag{2.11}$$

2.3.2 CV birth and Mass Transfer

When one of the stars in a binary overfills its Roche lobe, material from that star will enter the lobe of its companion via the L1 Lagrangian point. Mass transfer between binary components usually results in changes to the orbital separation and in turn the Roche lobe radii of the components, where mass ratio q plays an important role. Mass transfer occurs during two separate phases of evolution, i.e., on the path to becoming a CV, and during the CV phase (see e.g., Hellier 2001; Frank et al. 2002; Warner 1995).

During the pre-CV phase, the binary consists of two main sequence stars of unequal mass. The more massive star (primary) evolves off the main sequence first, expands in radius, and fills its Roche lobe, thus leading to the transfer of mass to the less massive companion (secondary). The centre of mass of the binary is closer to the primary, so the lost mass is moving away from this centre of mass, thereby increasing the angular momentum of the material. For the conservation of angular momentum to hold, the binary separation must decrease. A decreased separation leads to a reduction in the size of the Roche lobe (see Equation 2.6 and replace q with q^{-1} for the lobe radius of the primary), further increasing mass transfer. This positive feedback causes the unstable situation of accelerating mass transfer that eventually leads to an extended atmosphere around both stars, referred to as a common envelope phase. In this phase, matter is transferred too quickly for accretion onto the secondary such that not only are both Roche lobes filled but matter extends out to fill equipotential surfaces larger than that enclosed by the lobes. The common envelope that engulfs the binary acts to inhibit orbital motion, draining the binary of orbital angular momentum. The energy acquired by the envelope leads to its outward expulsion from the system, whilst also causing the binary orbit to shrink from $\sim 100\,\mathrm{R}_\odot$ to $\sim 1\,\mathrm{R}_\odot$. Once the envelope has dispersed we are left with either a semi-detached binary in the form of a cataclysmic variable or a detached binary, depending upon whether the orbital separation has shrunk enough for mass transfer from the secondary to the primary. In the former case, the primary will emerge as a white dwarf (typically of carbon-oxygen composition) accreting from its main sequence donor of lower mass.

In our newly formed CV, the transferred mass from the secondary (or donor) undergoes a loss in angular momentum as it moves close to the centre of mass of the system, which results in an increase in separation for the conservation of angular momentum to hold. Referring back to the Roche lobe radius equation (Equation 2.6), an increased separation causes an increase in the size of the donor's Roche lobe to the point where the donor can no longer fill it, and subsequently mass transfer ceases. To sustain mass transfer, the donor must either undergo further expansion due to nuclear evolution, or a mechanism must be present to shorten the orbital period and decrease the size of the lobe. The former is highly unlikely as the masses of secondaries are less than a solar mass, with slow evolution, the latter, however, can be achieved as a result of orbital angular momentum loss due to magnetic braking or gravitational radiation.

2.3.3 Disk Formation

The angular momentum loss sustained mass transfer produces an almost constant matter stream entering the primary Roche lobe through L1. Frank et al. (2002) provide a detailed account of the stream trajectory and disk formation which is summarised here. The stream is subject to Coriolis forces due to the rotating frame of reference. More specifically, the L1 point is orbiting perpendicular to the motion of the stream causing the stream to bypass the primary and swing around it whilst also being accelerated by the primary' gravitational potential well. The stream loops around the primary to intersect itself resulting in shocks that dissipate energy. Despite this, the stream cannot so easily rid itself of the angular momentum it had upon entering the lobe via L1. It therefore settles into the lowest energy orbit for its angular momentum, a circular one. The radius of this orbit corresponds to the specific angular momentum the stream had upon entering the lobe at L1, referred to as the circularisation radius, R_{circ}, (Equation 2.12), where G, P, a, b_1, and q are the gravitational constant, orbital period, binary separation, distance from the primary's centre to the inner Lagrange point L_1, and the mass ratio of the secondary to the primary, respectively. R_{circ} is always smaller than the lobe radius of the primary by a factor of 2–3.

$$\frac{R_{circ}}{a} = \left(\frac{4\pi^2}{GM_1P^2}\right) a^3 \left(\frac{b_1}{a}\right)^4 = (1+q)[0.500 - 0.227 \log q]^4 \tag{2.12}$$

Within the ring particle collision and shocks will heat the gas and cause energy to be radiated away. The energy comes from the gravitational potential so some of the gas responds by sinking deeper into the gravitational potential well of the primary, i.e. orbiting more closely. This inward-flowing matter entails a loss of angular momentum, though this can only occur with a transfer of angular momentum outwards to particles at larger radii in the ring, causing them to move to larger orbits. The situation is illustrated in Figure 2.8. The disk will spread until the inner edge transitions to the boundary layer above the surface of the primary. The outwards spread will be halted by tidal interactions with the donor.

2.3.4 Period Evolution via Angular Momentum Loss

The subsections above have established that once formed CVs must be subject to angular loss mechanisms for the donor to maintain contact with its Roche Lobe and mass transfer to continue. This results in an evolution of CVs from long periods ($\sim$ 6 hours or longer) to shorter periods followed by a rebound after a minimum orbital period is reached. Detailed accounts of hydrogen CV evolution can be found within Knigge et al. (2011), Hellier (2001), and Warner (1995). The following subsections aim to summarise these accounts.

Magnetic Braking

At the longest periods ($P_{orb} > 3$ hrs), magnetic braking (Verbunt & Zwaan, 1981; Spruit & Ritter, 1983) is believed to be the angular momentum loss mechanism that dominates. Magnetic braking involves the stellar wind and magnetic field of the CV donor. The magnetic field origin is poorly understood, though it is thought that a dynamo mechanism is established within the star when convection causes bubbles of gas into circular motions. The field is usually strongest in quickly rotating stars as in the red dwarfs of CVs that rotate at the orbital period of the system (hours). When the high-energy, ionised particles of the stellar wind approach the donor's magnetic field they cannot cross it, they are forced to spiral around the field lines, corotating with the secondary. The magnetic field accelerates the particles to the point where they can be flung from the system, taking away angular momentum from the donor. Since the donor is tidally locked to the primary, the angular momentum is taken away from the binary system itself, shrinking the orbit.

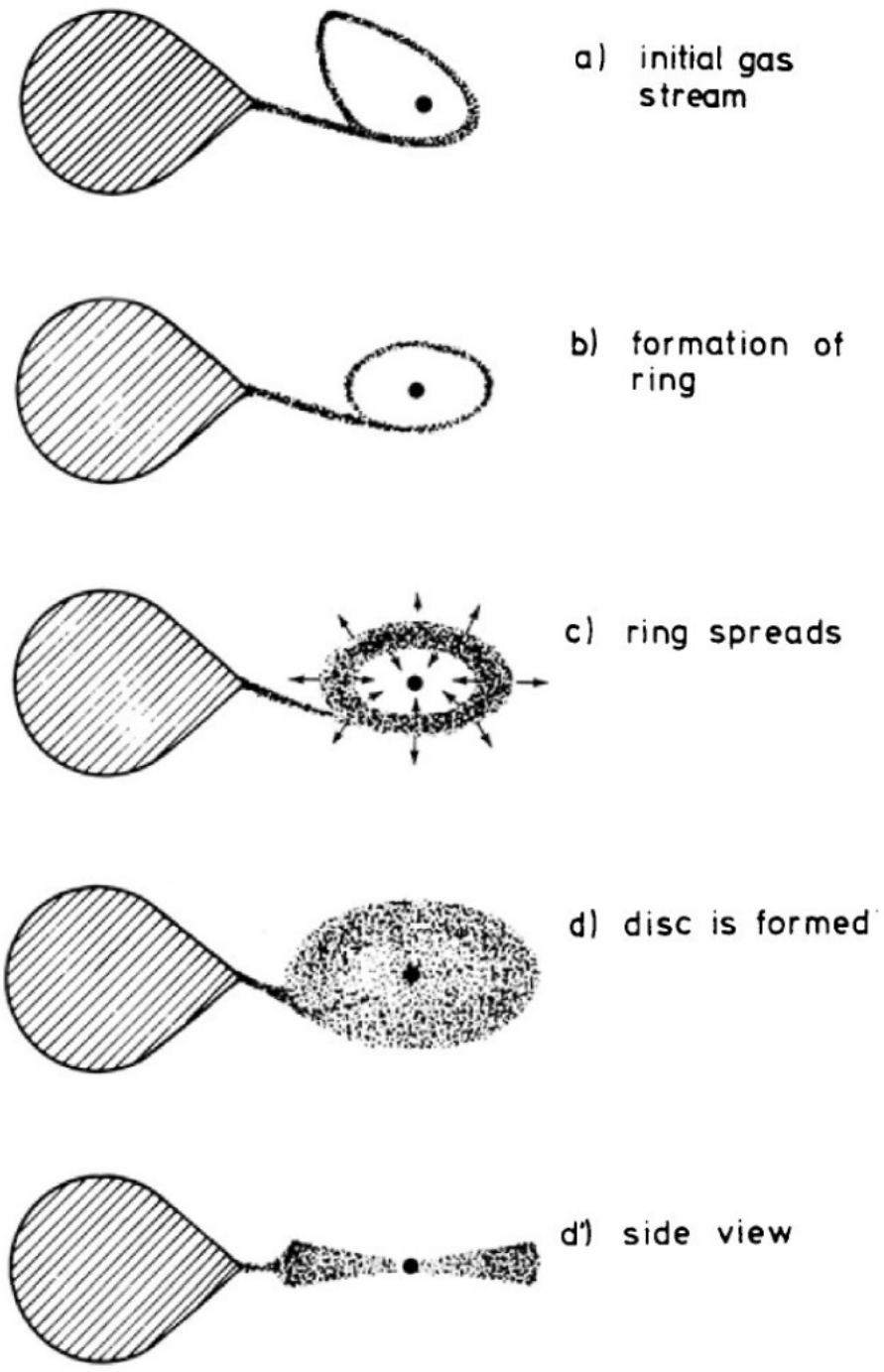


FIGURE 2.8: The figure above from Verbunt (1982) illustrates the formation of a ring and the evolution into a disk.

Period Gap: Disrupted magnetic braking

It emerged during the 1970s and early 1980s that there was a gap in the orbital period distribution of CVs between 2 and 3.12 hrs - referred to imaginatively as the 'period gap'. CVs in this period range were rarely observed (Figure 2.9). While the physical origin of this gap is still debated, current theories point to changes in the rate of angular momentum loss as the donor gradually loses mass to the WD (Knigge et al., 2011; Garraffo et al., 2018). With mass loss, the donor evolves to later and later spectral types. At the upper boundary of the period gap, the secondary will have become a fully convective dwarf star of spectral type M. This leads us to the theory of disrupted magnetic braking (Spruit & Ritter, 1983), currently the most popular explanation for the period gap. The magnetic dynamo that gives rise to magnetic braking is believed to occur at the boundary between the convective zone and the radiative interior. Therefore, when the star approaches the state of being fully convective the mechanism's effectiveness is greatly reduced or halted. Subsequently, the star shrinks to a radius appropriate for its mass, and loses contact with its Roche lobe, thereby halting mass transfer. The system becomes a detached binary; without the significant emission from the accretion disk or bright spot, the system becomes faint and less likely to be observable, hence the period gap. From here, gravitational wave radiation drives the system to shorter periods.

Re-establishment of mass transfer; Gravitational wave radiation

In the detached state, the continued evolution to shorter periods occurs via gravitational wave radiation (Paczyński, 1967; Paczynski & Sienkiewicz, 1981), which shrinks the Roche lobe of the donor. Gravitational wave radiation can be explained in the context of the theory of general relativity. Matter causes the fabric of spacetime to warp. Gravitational wave radiation refers to the ripples in this fabric caused by the acceleration of massive objects. These ripples (or waves) are more pronounced for objects with strong gravitational fields. In orbiting bodies, the waves propagate outwards at the speed of light, carrying away energy. In CVs, this results in angular momentum loss causing the orbit to shrink in a manner following Equation 2.13, where $d \ln J/dt$ is the time derivative of the natural logarithm of the total angular momentum of the binary, and P, G, c, M_1, and M_2 are the orbital period, gravitational constant, speed of light,

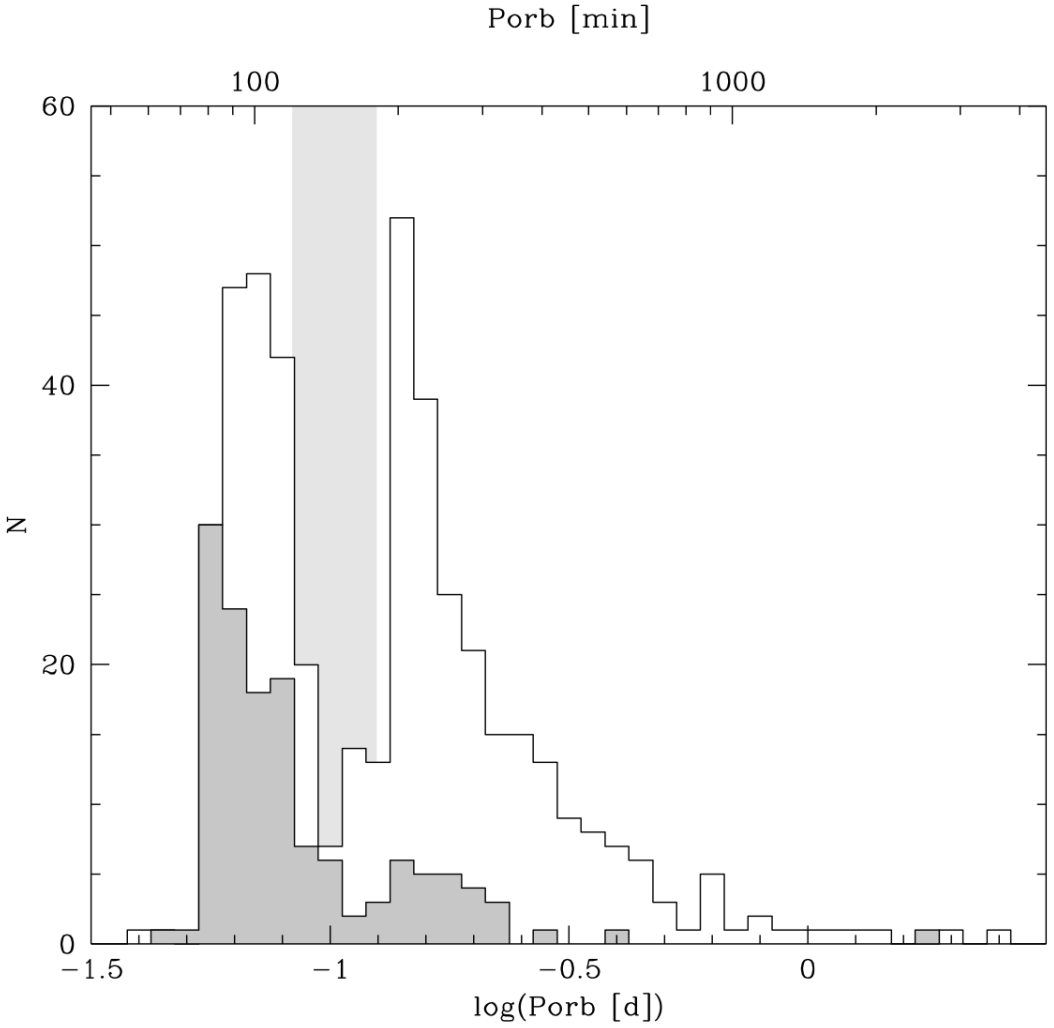


FIGURE 2.9: The orbital period distribution of 454 CVs from Ritter & Kolb (2003), V7.6, (white) and the distribution of 137 SDSS CVs from Gänsicke et al. (2009) (grey). The grey-shaded region represents the 2–3 h orbital period gap. The ultracompact ($P_{orb} < 65$ mins) hydrogen-deficient AM CVns (subsection 2.4.3) are excluded from the plot. Plot taken from Gänsicke et al. (2009).

and masses of the primary and secondary, respectively. Gravitational wave radiation is negligible for most binaries, though becomes significant for the shortest-period systems.

$$\frac{d\ln J}{dt} = -\frac{32}{5}\left(\frac{2\pi}{P}\right)^{8/3}\frac{G^{5/3}}{c^5}\frac{M_1 M_2}{(M_1+M_2)^{1/3}} \tag{2.13}$$

Continued evolution to shorter periods in the detached state eventually leads to a re-connection of the donor surface with its Roche lobe and a resumption of mass transfer at a period of ~ 2 hrs.

Period minimum and final long period evolution

The evolution to shorter periods via gravitational wave radiation continues until the donor mass becomes too low to sustain hydrogen burning, and the donor starts to become degenerate. Its response to further mass loss now becomes similar to that of white dwarfs, an increase in radius. Under non-degenerate conditions, mass loss to the high mass primary causes an expansion of the orbit, and an expansion of the Roche lobe that causes a cessation of mass transfer, only for the aforementioned angular momentum loss mechanisms to force the orbits to shrink and maintain sustained mass transfer. However, with this degeneracy, the donor expands with mass loss because such stars are no longer supported by the pressure of the gas but instead by the pressure of degenerate electrons, causing a reduction in mass to lead to an expansion of the radius (Lamers & Cassinelli, 1999). This allows the donor to maintain contact with its Roche lobe allowing for steady mass transfer despite an expanding orbit (and expanding lobe), without a requirement for angular momentum loss. Therefore, upon the onset of degeneracy, the orbital period increases with time. Such systems are often referred to as 'period bouncers'.

As a consequence, CV evolutionary theory predicts the existence of a minimal orbital period (e.g., Paczynski & Sienkiewicz 1983; Knigge et al. 2011) where, due to longer evolutionary timescales of shorter period systems, a spike in the number of CVs should be present around this minimum (or period spike). Theoretical estimates based on the standard model of CV evolution that invoke the angular momentum loss mechanisms described place the period minimum at around $P_{min} = 65 - 70$ mins (Kolb & Baraffe, 1999; Howell et al., 2001). However, the observed minimum period/period spike occurs at longer periods, with Gänsicke et al. (2009) finding a value in the range $80 < P_{orb} < 86$ mins with the aid of SDSS data thus implying an angular momentum loss in addition to gravitational wave radiation at short periods.

The evolution of the system from this point leads to the end of the system's time as a CV (Hellier, 2001). At the minimum period, the donor will resemble a brown dwarf star with a mass that will have dropped to only $\sim 0.06\,\mathrm{M}_{\odot}$. As the orbital period increases again the donor mass continues to decrease until at around 100 minutes, where its mass will only be around $0.02\,\mathrm{M}_{\odot}$. By this time the evolution has slowed and the mass transfer rate falls to levels that make the binary faint and difficult to detect and ultimately we end up with a Jupiter-like object orbiting a white dwarf. Alternatives to this endpoint depend upon the donor and/or white dwarf mass and accretion rate. For example, a

type Ia supernova may occur under conditions of a high mass white dwarf accreting at high rates (Hillman et al., 2016; Kato & Hachisu, 2012); or the CV may evolve into an ultracompact binary, where the donor transitions to transferring helium instead of hydrogen (Podsiadlowski et al., 2003).

System parameters through evolutionary stages

These evolutionary stages lead to an estimated donor mass evolution from $\sim 0.6\,M_\odot$ at $P_{orb} \sim 6$ hrs to below $\sim 0.04\,M_\odot$; and a mass transfer rate transitioning from $\sim 10^{-8}\,M_\odot yr^{-1}$ to $\sim 10^{-9}\,M_\odot yr^{-1}$ above the period gap and from a few times $10^{-10}\,M_\odot yr^{-1}$ to below $\sim 10^{-11}\,M_\odot yr^{-1}$ below the gap (Knigge et al., 2011). Knigge et al. (2011) also estimated that for a $0.6\,M_\odot$ donor where mass transfer was initiated at 6 hrs, the upper edge of the period gap is reached after $\sim 2.4 \times 10^8$ yr, evolution through the gap takes 0.4 Gyr, while the minimum period is reached after ~ 2.6 Gyr. Beyond the period minimum, the evolution is much slower such that longer evolutionary timescales are expected (Howell et al., 2001).

2.4 CV classification structure

The above discussion on emission components and the formation and evolution of CVs highlights the diversity of this transient class with systems lying somewhere on the continuum of orbital periods, mass transfer rates, mass ratios, and donor spectral types. The diversity is enhanced by the systems with a strongly magnetic white dwarf or a donor that is degenerate/semi-degenerate and helium-rich, which will be discussed shortly. The processes of mass transfer and orbital evolution under such a variety of conditions give rise to a diverse set of observable characteristics/phenomena witnessed in time series photometry and spectroscopy. Observational and subsequently derived physical properties of these systems allow for the division of CVs into several classes and sub-classes, which are often named after a prototype that is characteristic of its class/sub-class. The following classification structure is usually adopted.

- **Dwarf Nova**: Display semi-regular short-duration (days to weeks) outbursts. Main subclasses: U Gem, Z Cam, SU UMa;

- **Nova-like**: High accretion rate CVs; no outbursts; VY Scl subtype observationally distinct;
- **Nova**: Undergo thermonuclear eruptions with long (typically $10^4 - 10^5$ yr) recurrence times
- **AM CVn**: ultra-compact ($\sim 5 < P_{orb} < 65$ min) helium-rich CV
- **Polar**: Strongly magnetic WD; magnetically controlled accretion; accretion disk formation suppressed;
- **Intermediate Polar**: mildly magnetic WD; partial accretion disk formation.

In the following subsections, I will provide a more detailed description of these classes/subclasses. My descriptions of observational phenomena will largely be confined to the optical regime of the electromagnetic spectrum with an excursion to different wavelength regimes where necessary to explain the underlying physics. This will be in keeping with the wavelength regime that has been the focus of this research.

2.4.1 Dwarf novae

Dwarf novae (Warner, 1995; Hellier, 2001) show semi-regular brightenings, or outbursts that are typically 2-5 magnitudes in amplitude, recurring on timescales of days to years. Each outburst typically lasts a few days to a week, though longer durations are also seen. The values of amplitude, recurrence times, and duration are characteristic of a given system. Most outbursts have a rapid rise (several hours to a day) and a more gradual decline, though more symmetric profiles can also be seen. A variety of outburst profiles have been observed, as well as a variety of outbursting patterns, as shown in Figures 2.10, 2.11, and 2.12. To explain such variety as well as the characteristics that define dwarf nova subclasses, an understanding of the basic disk instability model is required. The model is widely believed to explain the majority of outburst phenomena.

2.4.1.1 Disk Instability Model

Current models of dwarf nova outbursts are based on the disk instability model (DIM), first proposed by Osaki (1974) and developed in the decades since (see Buat-Ménard

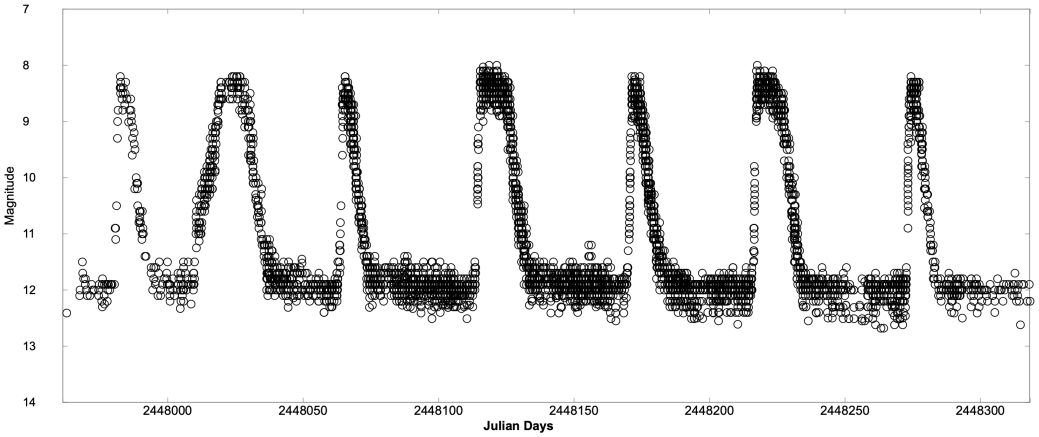


FIGURE 2.10: Light curve of unfiltered observations of SS Cyg from AAVSO spanning 1 year. While overall the system displays a semi-regular pattern of outbursts, both symmetric and non-symmetric dwarf nova outburst profiles may be seen. The rises are either fast ($\sim$ 2 days) or slow ($\sim$ 8 days), while the declines are all $\sim$ 8 days long. Plateaus are also present in 2 of the outbursts that last $\sim$ 10 days.

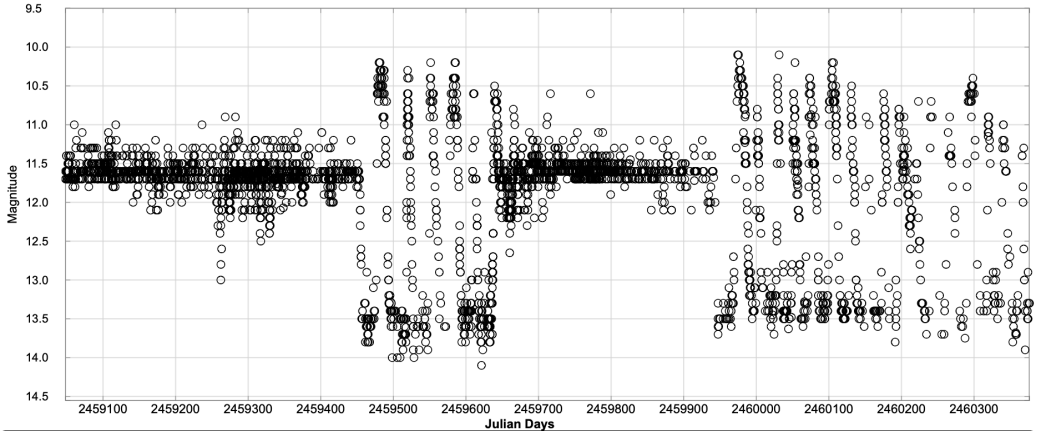


FIGURE 2.11: Light curve of unfiltered observations of Z Camelopardalis from AAVSO spanning 1 year. The system displays periods of rapid outbursts interspersed with periods of relatively constant brightness a few tenths of a a magnitude lower in brightness than max brightness (referred to as standstills).

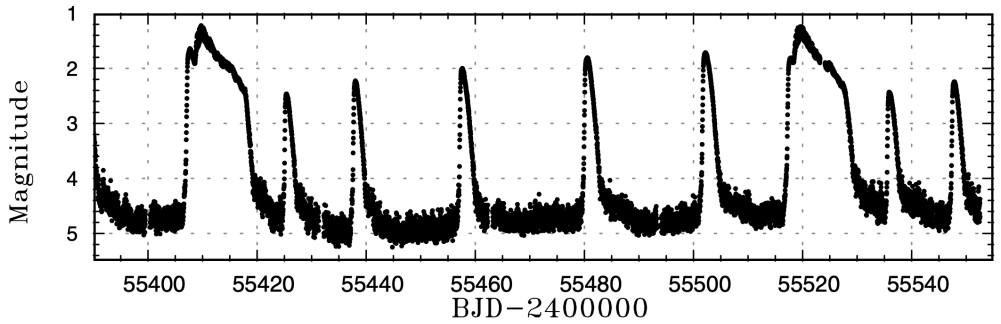


FIGURE 2.12: Kepler Light curve of V1504 Cyg showing two outbursts of longer duration and larger brightness with regular outbursts in between.

et al. 2001a and Hameury 2020 for detailed reviews and Hellier 2001 for a comprehensive summary). According to the DIM, for a disk in a low-viscosity state, the rate of mass transferred by the donor exceeds the rate of flow through the disk and onto the white dwarf, causing a build-up of matter in the disk. The temperature increases as a consequence, which leads to an increase in disk viscosity causing friction between adjacent annuli. In the Keplerian velocity structure, matter in outer annuli will be sped up (increasing its angular momentum), while inner annuli matter will be slowed (reducing its angular momentum). The outward travelling angular momentum causes the majority of matter to fall inwards, though a proportion travels outwards to transport the angular momentum. However, accretion disks are so diffuse that viscosity is too weak for such a process, a mechanism that acts as disk viscosity is required.

Viscosity

In 1973, Shakura & Sunyaev (1973) proposed that turbulence within accretion disks causes blobs of material to be transferred between adjacent annuli, transferring angular momentum, thus acting as disk viscosity. To model the turbulent viscosity in accretion disks, the alpha viscosity was introduced, defined as:

$$\nu = \alpha c_s H \tag{2.14}$$

where ν represents the viscosity, while the terms α, H, and c_s, refer (respectively) to a dimensionless parameter that defines the strength of the viscosity (number between 0 and 1), the disk scale height, and the local sound speed. The 'alpha viscosity' can be combined with equations for gas dynamics to generate models of accretion disks (or alpha disks). Such disks have heights much smaller than their radii; are slightly concave — flared at the outer edges; and have a mass negligible compared to the central WD. An alpha value of between $\sim 0.01 - 0.02$ is expected in a cold disk during quiescence, increasing to > 0.1 for a hot disk during a dwarf nova outburst (Frank et al., 2002). This alpha approximation though gives no clue as to the origin of the turbulence which creates the necessary viscosity.

Magnetic turbulence

Magnetohydrodynamic (MHD) instabilities are believed to be the origin of the turbulence (Hawley & Balbus, 1998). Consider a weak vertical (perpendicular) magnetic field

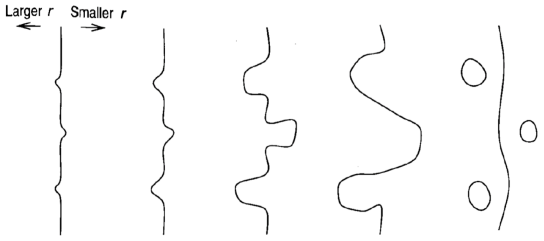


FIGURE 2.13: The figure from Hellier (2001) illustrates the growth of the Balbus Hawley instability that is sequenced from left to right. Small kinks in the field (perpendicular to the plane of the accretion disk) are amplified by differential matter flow. This increases the strength of the field until a reconnection occurs to dissipate energy. Since the ionised material follows the field lines, bubbles of gas are transported to different radii.

within the disk; free charges within the ionised disk matter may flow along field lines but cannot easily cross them, while field lines can be stretched by the motion of ionised matter. A small radial kink in an otherwise vertical field will be subject to two opposing forces: magnetic tension will act to straighten out the kink while the differential velocity of matter in annuli on either side of the field line acts to stretch out the kink. Matter in the annulus immediately inside of the field line (at a smaller radius) moves faster than that within the annulus immediately outside the line. This stretches the kink in the line. However, inner annulus matter is slowed by the field, loses angular momentum, and falls to shorter radii, while the outer annulus matter is sped up by the field, increases in angular momentum, and moves to larger radii. This stretches the magnetic field further (making it stronger), thus enhancing the matter redistribution (or magnetic turbulence). This process is referred to as the Balbus-Hawley instability that provides the necessary viscosity. This situation is illustrated in Figure 2.13.

S Curve

When the accretion disk is hot, the material is readily ionised allowing particles to interact with the magnetic field instigating the Balbus-Hawley instability and an outburst state. The instability shuts down in cool, neutral disks, corresponding to a dwarf nova in quiescence. The alternating between neutral (cool) and ionised states (hot) is required for dwarf nova outbursts. The S Curve can be used to describe these transitions in terms of a cycle describing a dwarf nova outburst. The S curve is a plot of disk surface density Σ versus surface temperature T_{surf} (Figure 2.14).

Quiescence corresponds to point A in Figure 2.14 - a cold, neutral, low-viscosity disc. As matter piles up in the disk, the surface density increases. This increases particle

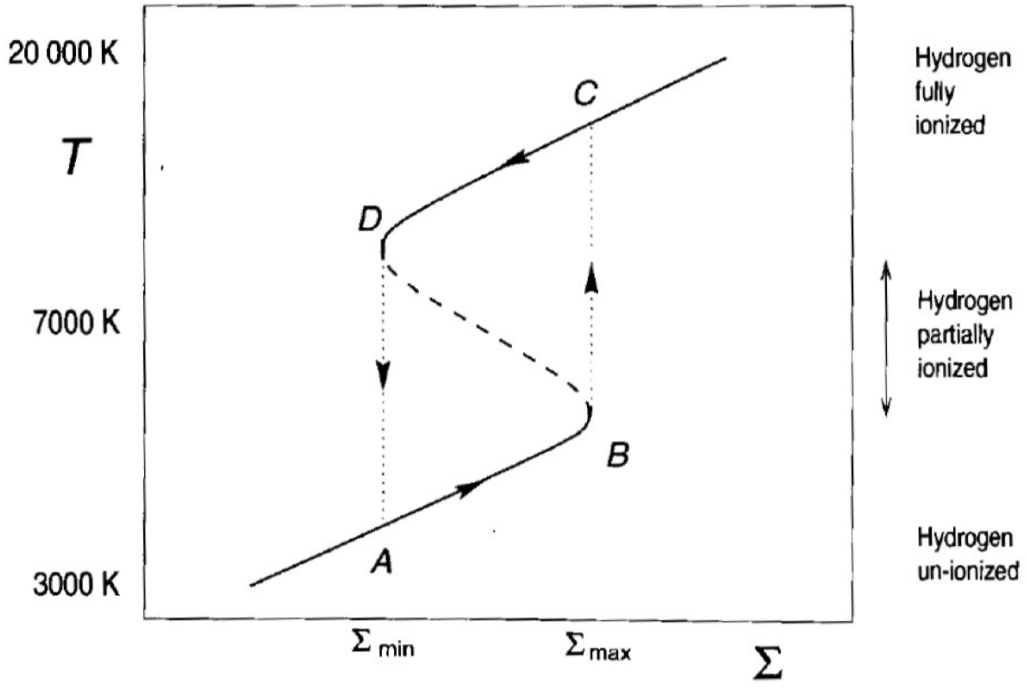


FIGURE 2.14: The figure from Hellier (2001) shows the dwarf nova cycle plotted as disk surface temperature T as a function of disk surface density Σ_{surf}. The S-curve forces the disk to follow the cycle from $A \Rightarrow B \Rightarrow C \Rightarrow D \Rightarrow A$.

interaction (or viscosity), which in turn raises the hydrogen-rich disk temperature to around 7000 K. The temperature is sufficient to partially ionise the gas, boosting its opacity. Free electrons can then combine with neutral hydrogen to create H^- ions that are particularly effective at absorbing photons. Any further increase in temperature will now lead to a large increase in opacity. An increase in surface density drives this increase in temperature and thus opacity and in turn a runaway temperature increase. The timescale for heating is far shorter than that required for the viscous exchange of matter such that the surface density remains roughly constant through this phase as the temperature rises until the hydrogen is completely ionised, point C. This new equilibrium state of higher luminosity due to increased temperature is maintained by the higher inward flow of material driven by the increased viscosity of the ionised material.

This state is temporary as the rate of flow of matter through the disk now exceeds the rate of matter arriving from the donor. Therefore, the surface density drops as does the surface temperature resulting in the system moving to point D - the disk returns to a partially ionised state. The high opacity now traps heat in the mid-plane rather than the surface of the disc due to the decreased surface density. The disc surface temperature

therefore plummets, the ions recombine, the viscosity drops, and the disc returns to a cold quiescent state, point A. This cycle repeats itself on timescales largely governed by the time taken for matter to build in the disk again - the rate of mass transfer. A mathematical prescription of this process is provided within Hellier (2001).

Heating and cooling waves

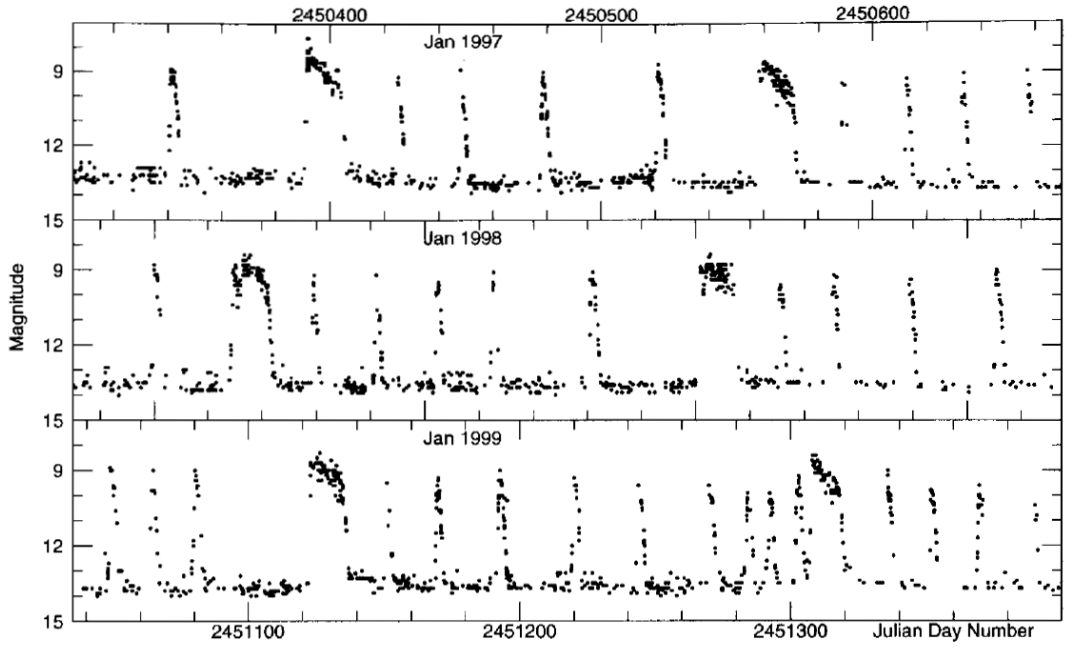
The transition from quiescence to outburst state described above does not happen across the whole disc in unison, rather, the process usually begins at a particular annulus. This would be the annulus that first reaches point B in Figure 2.14 corresponding to a critical surface density Σ_{max}. The higher viscosity spreads hot material from this annulus to adjacent annuli generating thermal instability throughout the disc in a domino effect. This 'heating wave' spreads throughout the disc driving the system into outburst. Should Σ_{max} be met first in the inner disk, the heating wave propagates outwards, creating an 'inside-out' outburst (or type B outburst), whereas, if Σ_{max} is met in the outer disk first, the heating wave propagates inwards, creating an 'outside-in' outburst (Type A). Type A outbursts usually result in an asymmetric outburst profile, with a rise more rapid than the decline. This occurs because viscosity causes more material to flow inward than outward, and annuli at larger radii have a higher surface density and volume of matter compared to those at smaller radii. Combined, these factors enable the inward-running heating wave to quickly overwhelm the next annulus, propagating rapidly through the disk. Type B outbursts, on the other hand, tend to have a symmetric profile with a slower rise rate. This is due to the same factors, which impede the outward-running heating wave's progress. The enhanced accretion rate drains material until Σ is reduced to Σ_{min} at some annulus, usually in the outer disk. This annulus falls out of outburst instigating a cooling wave spreading inwards and placing the system into quiescence. Outside-in outbursts occur if the time scale for matter accumulation at the outer disc edge is shorter than the timescale for the viscous diffusion of matter to short radii, and are therefore associated with high mass transfer rates. Type B outbursts are associated with low mass transfer rates for the inverse of the reasons for type A outbursts. The significance of the annuli at which the outburst is instigated is believed to determine the shape of a dwarf nova outburst. This is explored in the following subsections on the different classes of dwarf nova.

2.4.1.2 Dwarf Nova: U Geminorum (U Gem)

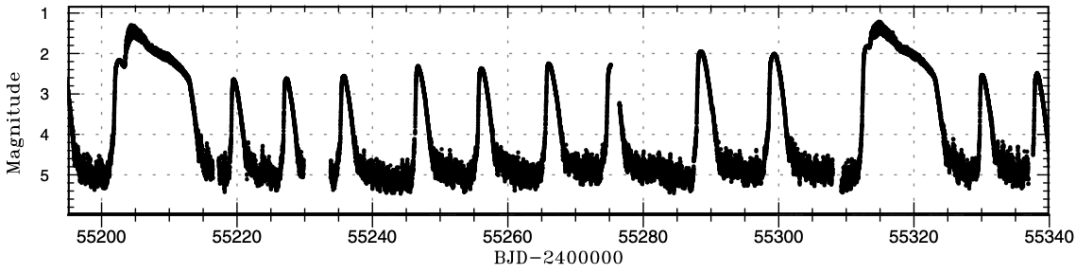
U Gem dwarf novae (e.g., Figure 2.10) only display semi-regular dwarf nova outbursts. These systems usually lie above the period gap (> 3.12 hrs) with mass transfer rates $\geq 10^{-9} M_{\odot} yr^{-1}$. Estimates of typical outburst amplitudes, durations, and cycle lengths (time between successive outbursts) of individual dwarf nova systems have been measured by Otulakowska-Hypka et al. (2016) based on available light curve data. Otulakowska-Hypka et al. (2016) estimated the following properties for U Gem systems in their dataset: optical amplitudes range from 1.25 to 6 magnitudes (although high orbital inclinations, where the disk is viewed closer to edge-on, may lead to underestimations); recurrence periods vary between 5 days and 250 days; and the duration of outbursts ranges from 2–3 days to 23 days.

2.4.1.3 Dwarf Nova: SU Ursae Majoris (SU UMa)

The SU UMa subclass of dwarf novae (see e.g., Warner 1995; Hellier 2001; Osaki & Kato 2013) are typically short orbital period systems residing below the period gap. In addition to the 'normal' outbursts exhibited by U Gem stars, they also exhibit especially long outbursts called "superoutbursts" that are ~ 1 mag brighter with a typical duration of about two weeks in contrast to the several-day durations of the shorter normal outbursts (Figure 2.15). The supercycle length (time between successive superoutbursts) of any given system is always longer than its cycle length. Superoutbursts coincide with the peak of an underlying modulation of the light curve, called superhumps, whose amplitudes are on average a few tenths of a magnitude with periods a few percent longer than the orbital period (Smak, 2010). The precession of an ellipticity in the disk, brought about by tidal interactions with the donor star, as described by the tidal-thermal instability model (Whitehurst, 1988; Osaki, 1989), is the commonly accepted explanation for superoutbursts and superhumps. Typical properties of individual SU UMa stars measured by Otulakowska-Hypka et al. (2016) are summarised as follows: normal outburst amplitudes within the range of 1 and 6 magnitudes in the optical; cycle lengths within the range of 4 and 398 days; superoutburst amplitudes between 1.5 and 7 magnitudes; supercycle lengths between 50 and $\sim$2000 days; and superoutburst durations around 10 to 30 days.



(A) VW HYi



(B) V1504 Cyg

FIGURE 2.15: (A) Light curve of the SU UMa type dwarf nova VW HYi showing both normal and superoutbursts. Data taken from the Royal Astronomical Society of New Zealand. (B) Kepler light curve of V1504 Cyg (Osaki & Kato, 2013).

SU UMa subclass: ER UMa

SU UMa systems with very short supercycle lengths are referred to as ER UMa stars. They have extremely high outburst frequencies (3-4 day outburst cycle), supercycle lengths between 19 and 48 days, and the presence of superhumps (Kato et al., 2013). Based on only 11 systems, Smak (2010) found normal outburst amplitudes lie between 1 and 3 magnitudes with recurrence times between 2.8 and 5.6 days; superoutburst amplitudes between 2 and 4 magnitudes with recurrence times between 17.8 and 80 days; normal outburst amplitudes lie between 1.5 and 3.2 magnitudes; superoutbursts between 2 and 4 magnitudes; outbursts duration of normal outburst within 1 and 4 days; and superoutburst duration between 10 and 30 days.

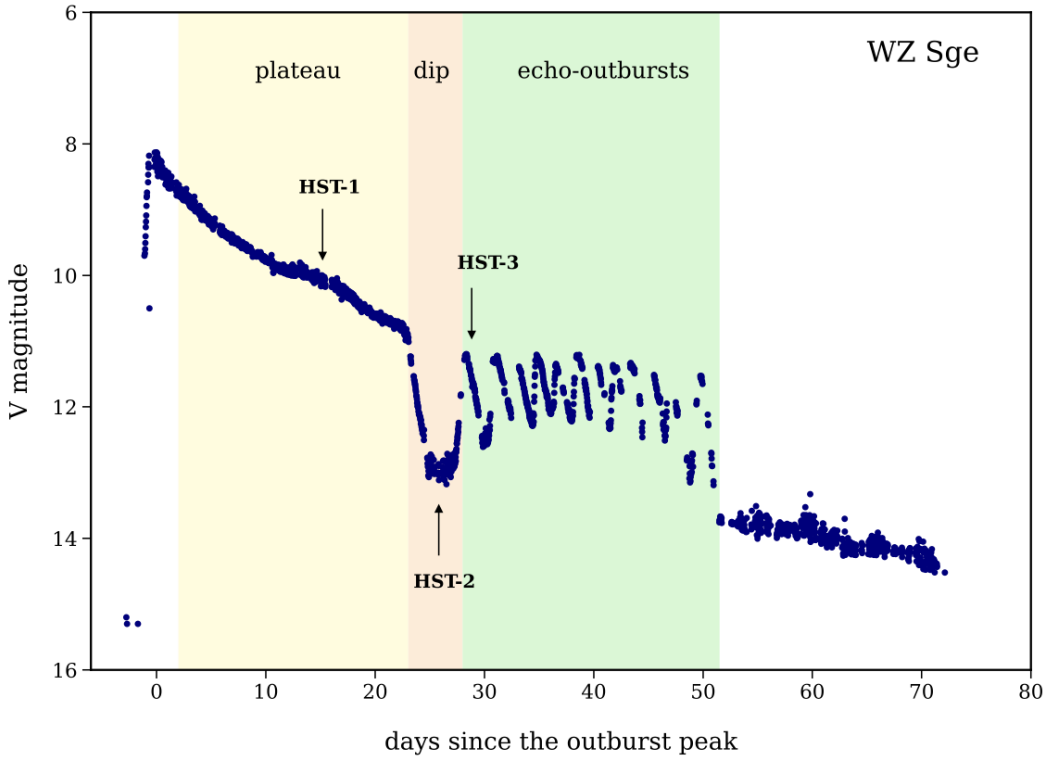


FIGURE 2.16: VSNET optical light curve of WZ Sge's 2001 superoutburst (Georganti et al., 2022). Three different regions are highlighted: the plateau (yellow), the dip (orange), and the echo-outburst phase (green).

SU UMa subclass: WZ Sagittae

The WZ Sge subclass (see Kato 2015 for an extensive review) have supercycle lengths of order years, where observations reveal a median of 11.5 years and a majority below 40 years. Other defining characteristics are: the absence of normal outbursts; large amplitude outbursts (typically $\sim$ 8 magnitudes) at least greater than 6 magnitudes; slow declines from superoutbursts (weeks as opposed to days); in addition to 'normal' superhumps, double-wave modulations at the orbital period that last at least several days during the early stage of the outburst - referred to early superhumps; and the presence of multiple rebrightenings on the fading tail of the superoutbursts have been considered supporting evidence. A light curve of WZ Sge is provided in Figure 2.16. Most systems have orbital periods shorter than $\sim$86 minutes and comprise the 'period minimum spike' distribution of CVs (Gänsicke et al., 2009) between 80 and 86 minutes. These have very low accretion activity owing to brown dwarf donors.

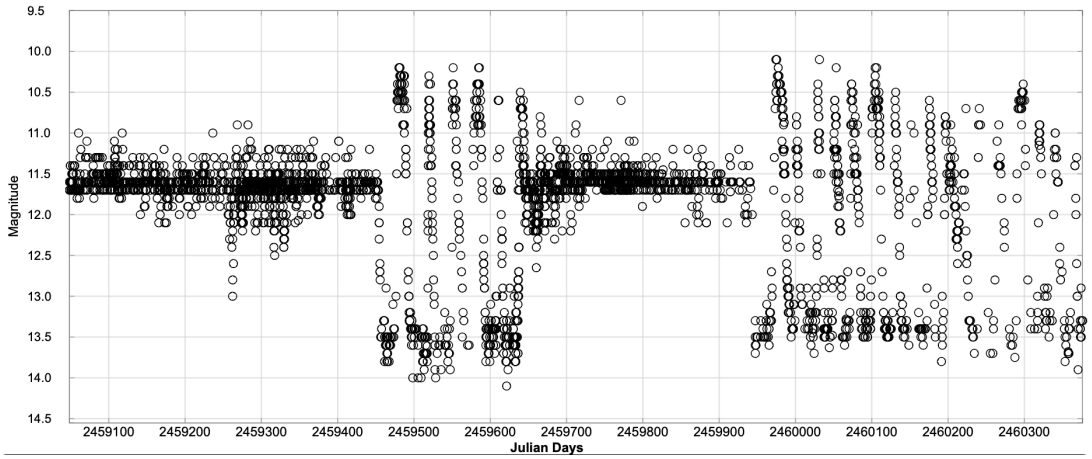


FIGURE 2.17: Light curve of Z Cam formed with AAVSO visual band data

2.4.1.4 Dwarf Nova: Z Camelopardalis (Z Cam)

Z Cam-type dwarf novae (Simonsen et al., 2014) are characterised by periods of outburst activity with short cycle lengths, spending little time in quiescence (Figure 2.17). These periods are interrupted by standstills (instigated by an outburst), where the system maintains a brightness between outburst maximum and quiescent minimum for days to weeks. Standstills end by returning to quiescence on a timescale of order the decline rate of the outbursts. Otulakowska-Hypka et al. (2016) and Simonsen et al. (2014) find typical cycle lengths are between ∼5 and ∼56 days; outburst amplitudes are generally lower than SU UMa and U Gem systems, within a range of 2.3 and 4.9 magnitudes; and outburst durations between 2.5 and 25 days are seen, with a majority shorter than 15 days. Furthermore, the orbital period distribution of Z Cam systems shows they reside above the period gap with a range of 3.1–8.4 hours (Simonsen et al., 2014), the average is around 5.3 hours.

To explain the standstill phenomenon, one may refer to a critical mass transfer rate (see e.g., Dubus et al. 2018) that is a function of the orbital period. Above this critical rate, the accretion disk is hot and stable (in a high state). In contrast, below this rate, the disk is cool and unstable to dwarf nova outbursts whereby a low state is present most of the time but interrupted by brief excursions into a high state (dwarf nova outburst). Figure 2.18 shows the stability criterion marked by the red line. Z Cams tend to possess mass transfer rates higher than other dwarf novae and are expected to lie close to this stability limit (Meyer & Meyer-Hofmeister, 1983; Buat-Ménard et al., 2001b). The standstills are thought to be due to fluctuations of the mass transfer rate of 10-30% bringing it very

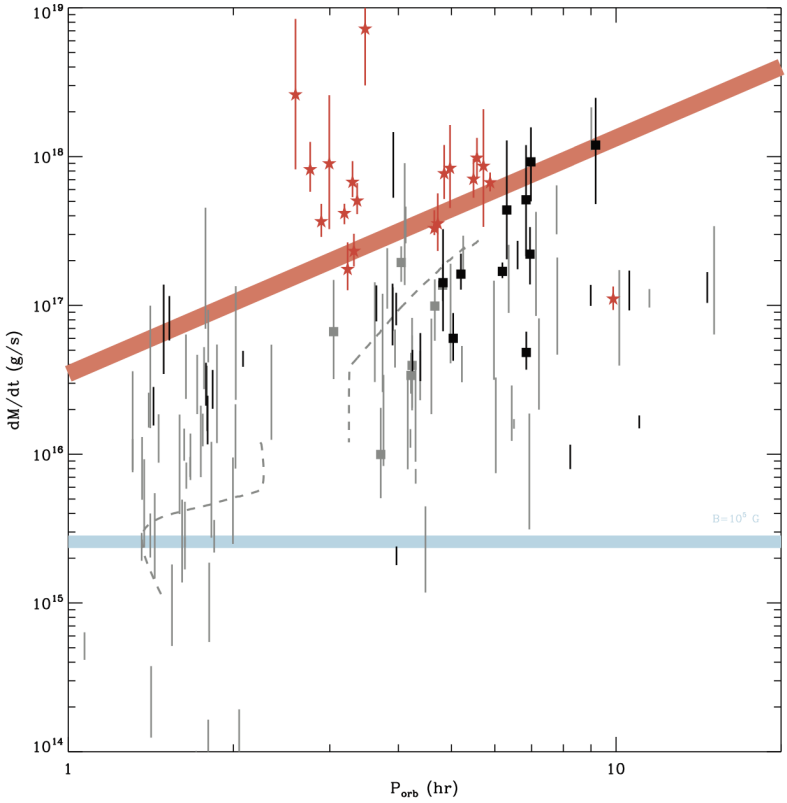


FIGURE 2.18: Mass transfer rates of CVs compared to stability criterion. Systems above the (red) upper solid line are hot and stable while systems below the lower (blue) line indicate cold, stable disks. Square symbols indicate Z Cam systems; (red) stars indicate nova-likes (discussed shortly). Z Cams tend to lie close to the red line (Dubus et al., 2018).

close to or above the stability limit; the origin of the mass transfer rate variations is uncertain.

2.4.2 Nova-likes

Nova-like systems (Warner, 1995; Meyer & Meyer-Hofmeister, 1984; King & Cannizzo, 1998) are characterised by high mass transfer rates and are typically found above the period gap. Nova-likes, U Gems and Z Cams seem to overlap in orbital period ranges, implying that mass transfer rates can vary significantly for different systems with similar orbital periods. Apart from the VY Scuptoris subclass, nova-like systems show little photometric variability in comparison to all other CV types. They typically exhibit no outburst activity, varying only slightly about their mean level unless the system inclination is sufficiently high to generate an eclipse. For a given orbital period their mass transfer rates are sufficiently high to allow the disk to be maintained in a high

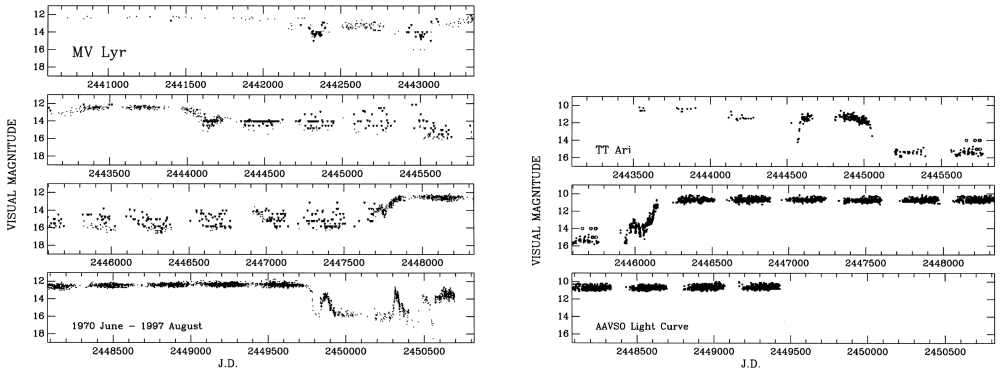
state - or permanent outburst state (see Figure 2.18 for the location of these systems with respect to the critical mass transfer rate). According to the definition of Warner (1995), nova-likes include all non-eruptive CVs, so conceivably they could include pre and post-novae (see subsection 2.4.5) or Z Cams in standstill, where our observational timeline is too short to reveal any outburst. Magnetic CV types (Section 2.4.4) are sometimes included with nova-likes, but are usually treated separately.

There exist subtypes of nova-likes: SW Sextantis stars have the highest mass transfer rates, and along with the UX Ursae Majoris and RW Trianguli subclasses may only be spectroscopically distinguished. VY Scultoris (VY Scl) systems are the only subclass photometrically distinguishable from the others. Therefore, the remainder of this subsection is focused on them due to the focus of my research on automated classification based on photometric variability.

The VY Scl subclass appears mostly above the period gap, at the 3–4 hours range. They display pronounced low states with depths up to (and occasionally exceeding) 5 magnitudes that interrupt high states at irregular intervals (Honeycutt & Kafka, 2004). The transition from high to low state usually occurs on timescales of weeks to months. The low states are believed to be caused by a temporary reduction or cessation of mass transfer from the donor. While low states also occur in strongly magnetic CVs, I restrict the use of 'VY Scl' to non-magnetic CVs. A possible cause of such states, put forward by Livio & Pringle (1994) is the migration of a star spot or multiple star spots on the donor to a region directly underneath the inner Lagrangian point (L1) that connects the binary components; this theory is currently believed the most viable (Honeycutt & Kafka, 2004). Figure 2.19 shows the light curves of VY Scl systems MV Lyr and TT Ari (Leach et al., 1999). Both display clear low state excursions; in MV Lyr, the low state is ∼4–5 magnitudes lower than the high state, while in TT Ari the low state is ∼6 magnitudes fainter than the high state.

2.4.3 AM Canum Venaticorum (AM CVn)

The AM Canum Venaticorum stars (Solheim, 2010; Levitan et al., 2015) are ultra-short period (5-65 minutes) binaries where the donor star is of mostly helium composition. They remain rare, with 56 known systems reported during the last review (Ramsay et al., 2018), and recent discoveries (e.g., van Roestel et al. 2021, 2022) increasing the



(A) MV Lyr (B) TT Ari

FIGURE 2.19: Light curves of VY Scl type nova-likes (A) MV Lyr and (B) TT Ari.

number of published systems to approximately 80 (a small fraction of the thousands of CVs currently discovered, e.g., Breedt et al. 2014). The donor composition is uncertain but theorised to be either another white dwarf of lower mass, a stripped semi-degenerate helium star, or an evolved CV donor. They are characterised by their blue colour, due to the accreting WD dominating the flux contribution over an extremely low mass donor (within Gaia DR3; Gaia-Collaboration et al. 2022) the BP-RP colour is typically less than 0.6). Strong helium emission and the absence of hydrogen within their spectra are a key property.

2.4.3.1 Formation and evolution

The formation of AM CVn differs slightly from that of hydrogen CVs. Three possible channels exist for the evolution of the donor in an AM CVn, with the relative importance of each rather uncertain. One possibility is another WD of lower mass, rich in helium (Paczyński, 1967; Faulkner et al., 1972). Another is a helium star donor (Iben & Tutukov, 1987). The final scenario is the remnant of a low-mass main sequence star that has lost most of its hydrogen during its life as an ordinary CV (Podsiadlowski et al., 2003). Initially, all involve a close main sequence binary going through one or more common envelope phases as the stars evolve off the main sequence. Figure 2.20 is a schematic representation of possible pathways to becoming an AM CVn (Solheim, 2010).

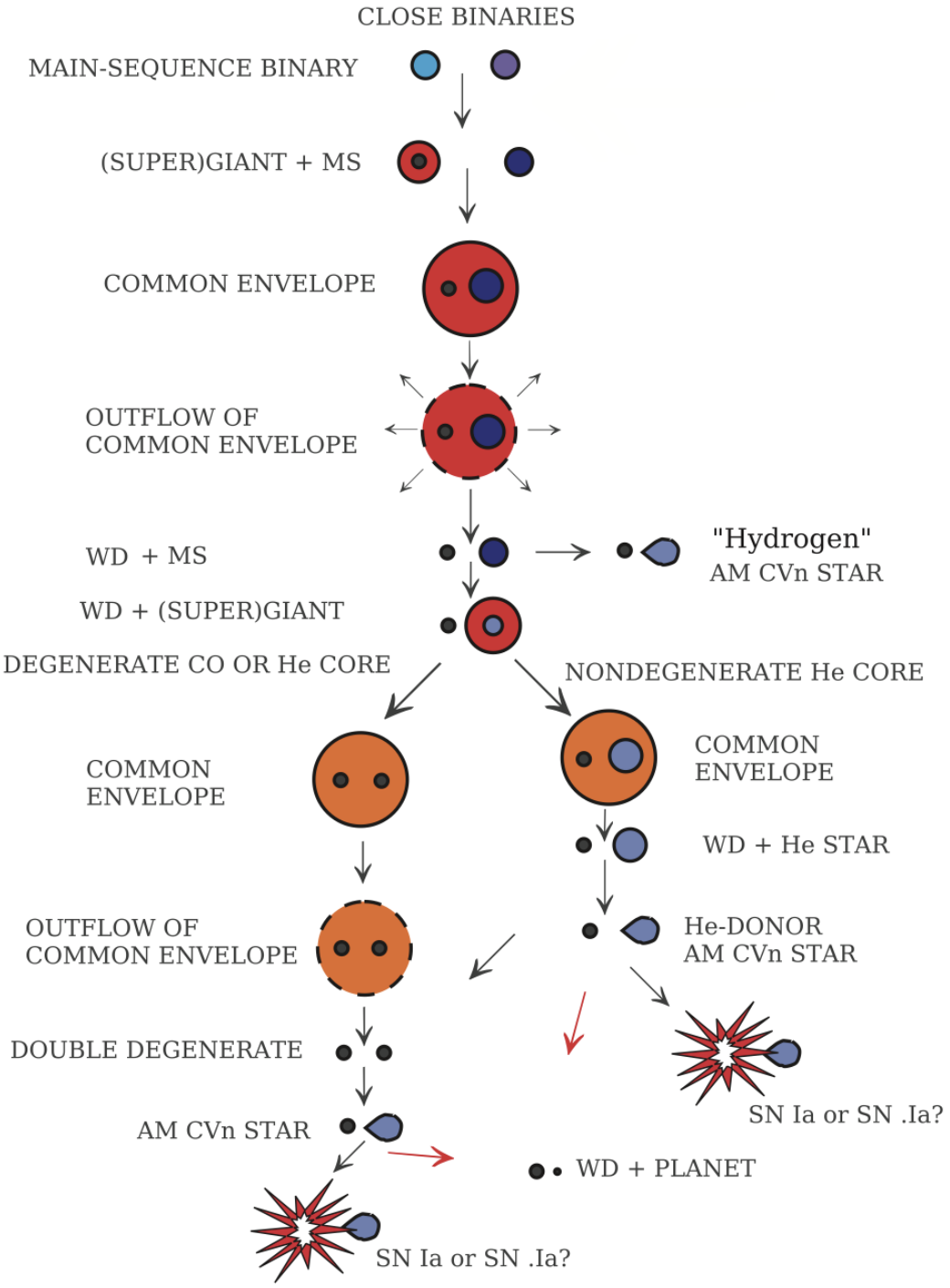


FIGURE 2.20: The possible evolutionary pathways in AM CVn stars from close binaries to supernova explosions or a cooling white dwarf with a companion (Solheim, 2010).

We begin with two main sequence stars, each with masses low enough to develop WD cores. The more massive star (primary) will be the first to evolve into a giant (or supergiant), and should the binary separation be short enough, the first common envelope (CE) phase will be established. This phase draws the binary closer together. Following the CE phase, the primary will emerge as a white dwarf with a helium or carbon core. It is from here we explore the three pathways.

For a white dwarf donor, the eventual donor mass after the first CE phase should be $M_2 \leq 2.3\,\mathrm{M}_\odot$. Then following its evolution to a giant and through the second CE phase it will emerge as a helium white dwarf and make up the subset of double white dwarf AM CVns (Paczyński, 1967; Faulkner et al., 1972); this is shown as the left branch of the Figure 2.20. Once gravitational wave (GW) radiation has shrunk the orbits to a minimum of around 5 minutes, the less massive WD will begin to transfer mass via Roche lobe overflow. Shortly after the start of mass transfer, the orbital evolution will reverse, causing the binary separation to increase. The binary is most likely observed after the orbital period minimum.

The helium star channel is much the same as for the double white-dwarf channel except that the secondary emerging from the first CE phase and causing the second CE phase is more massive $2.3\,\mathrm{M}_\odot \leq M_2 \leq 5\,\mathrm{M}_\odot$, and goes on to become a helium star donor (Savonije et al., 1986; Iben & Tutukov, 1987) rather than another WD. This is shown as the right branch of Figure 2.20. Donor helium burning begins shortly after the second CE phase. Mass transfer begins once GW radiation has brought the systems close enough together, usually around $P_{orb} \sim 10$ minutes, soon after which the orbital period increases. The helium star will become increasingly degenerate as it evolves to longer periods.

For an evolved donor (Podsiadlowski et al., 2003) a second CE phase (involving the secondary) does not occur, but the donor fills its Roche lobe and begins mass transfer near the end of its main sequence lifetime (terminal age main sequence or TAMS). Therefore, they appear as CVs with evolved donors in their early evolution. Magnetic braking brings the system to ultrashort periods dependent on when mass transfer begins in relation to the TAMS. Systems which evolve this way may be observed either before or after the minimum period is reached (between 5 and 70 minutes). These donors are

initially assumed to have some hydrogen on their surface, though become increasingly degenerate and helium-rich.

Mass transfer in AM CVn systems is expected to continue until one component becomes a dark sub-stellar object, though a type Ia or .Ia supernova may be another possible endpoint should the accreting WD reach the Chandrasekhar mass.

2.4.3.2 Photometric behaviour

As with hydrogen CVs, the photometric behaviour of AM CVns tends to be governed by the mass transfer rate, which is a strong function of the orbital period (Cannizzo & Nelemans, 2015; Solheim, 2010; van Roestel et al., 2021). At orbital periods, $P_{orb} < 10$ minutes, high mass transfer rates are present, and the accretion stream directly impacts the accreting white dwarf — so no accretion disk is present. They can be detected through their X-ray emission modulating at the orbital period (e.g., in HM Cnc Roelofs et al. 2010). An accretion disk may form for systems with slightly longer periods ($\sim 10 < P_{orb} <\sim 22$ minutes). Mass transfer rates are high enough to sustain the disk in a constant high state (Green et al., 2018). At intermediate periods ($22 \lesssim P_{orb} \lesssim 45$ minutes), the accretion disk is unstable to the kind of outbursts and superoutbursts present in hydrogen CVs (van Roestel et al., 2021). As the orbital period increases within this range, outburst recurrence times increase exponentially, while the luminosity of the disk decreases (Levitan et al., 2015; Nelemans et al., 2004). At $P_{orb} > 45$ minutes the mass transfer rate is so low that the accretion disk is cool, optically thin, and outbursts are rare.

Kato & Kojiguchi (2021) defined a set of photometric variability criteria/characteristics by which one may identify a system as an outbursting AM CVn (these criteria are by no means strict).

- Rapid fading in any part of the light curve (more than 1.5 magnitudes/day);
- Short duration superoutbursts of usually 5-6 days (in hydrogen CVs this is typically greater than 10 days);
- superoutburst amplitudes of 4-6 magnitudes (generally lower than in hydrogen CVs due to smaller disk);

- Double superoutburst where a rapid fading after the first outburst is a signature of an AM CVn and worth observing for a second superoutburst;
- Emergence of superhumps following second superoutburst if present;
- long fading tail after superoutbursts (100-200 days);
- multiple rebrightening events on the fading tail;
- rebrightenings show a rapid decline with a rate of more than 2 mags per day;
- lack of red excess during this fading tail;
- strong UV excess, or blue colour in quiescence;
- faint absolute magnitude (significantly fainter than +4) of outburst is a sign of an AM CVn-type superoutburst.

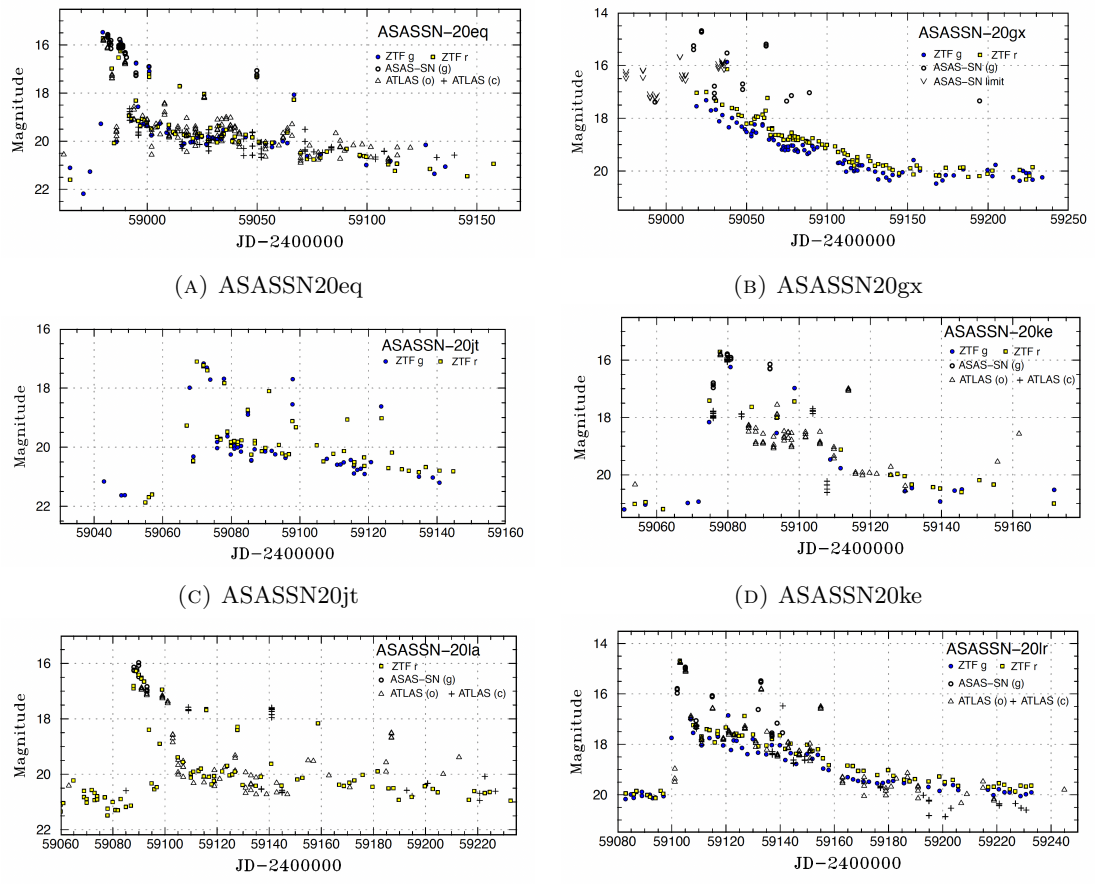
Figure 2.21 shows a selection of AM CVn outbursts for a variety of systems.

2.4.4 Magnetic CVs

2.4.4.1 Magnetically controlled accretion

In around $\sim 25\%$ of CVs, the magnetic field of the WD is strong enough to affect the motion of the charged particles within the accretion stream arriving from the donor. As described by Hellier (2001), at large distances from the WD the kinetic energy of the stream exceeds that associated with its interaction with the field. Matter within the stream will continue its trajectory unaffected by the field (dragging the field along with it). Close to the WD, the energy of the matter-field interaction exceeds the kinetic energy of the matter stream. The charged particles will be diverted out of the orbital plane and be forced to spiral around and move along the field lines to the white dwarf surface.

Since the strength of the magnetic field declines with distance from the white dwarf, there is a transition region/radius between the two scenarios. This transition may be referred to as the magnetospheric boundary within which matter is forced to follow the field lines in corotation with the WD and eventually impact the WD surface at or



(A) ASASSN20eq

(B) ASASSN20gx

(C) ASASSN20jt

(D) ASASSN20ke

(E) ASASSN20la

(F) ASASSN20lr

FIGURE 2.21: Light curve of several AM CVns during outburst (Kato & Kojiguchi, 2021).

near one or both of the WD poles. The strength of the boundary is aided by magnetic screening, where it induces a current in the plasma that counteracts the effect of the field, screening the field from matter further out that is unaffected by the field.

The spin period of the WD tends to adjust itself to match the circular Keplerian velocity of the matter just outside the magnetosphere. This marks an equilibrium situation where there will be no large jump in velocity at the boundary. As a consequence, the lowest field WDs have the smallest magnetospheres and the shortest spin periods, while the highest field WDs will have the largest magnetospheres with spin periods matching the slowest moving regions of the binary. Since diverting the stream out of the plane of the orbit requires energy, the WD magnetic axis will tend to align itself with the direction from which the stream is coming. Polarisation of the light from polars can be used to deduce the orientation of the magnetic axis with respect to the spin axis.

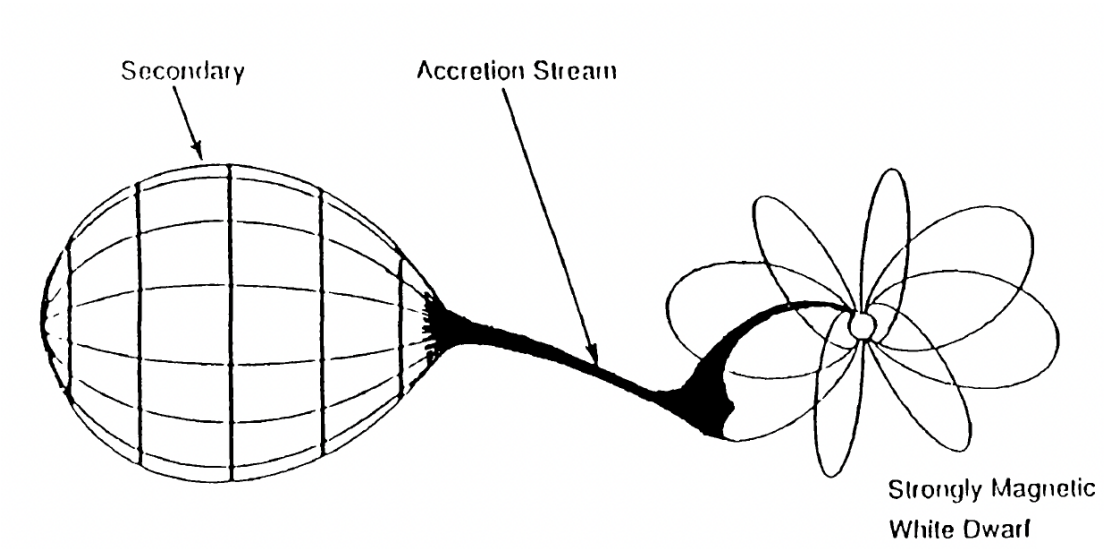


FIGURE 2.22: Schematic of a Polar CV from Cropper (1990)

2.4.4.2 Polars

Physical properties

Where the WD possesses a strong magnetic field, $B > 10MG$, the radius of the magnetospheric boundary is larger than the circularisation radius - the radius of a Keplerian orbit with the same angular momentum as the accretion stream at the L1 point (Frank et al., 2002). Consequently, the formation of an accretion disk is completely inhibited, instead, the accretion stream is directed out of the orbital plane and follows the magnetic field lines directly onto one or both of the WD's magnetic poles. Referred to as polars, or AM Herculis stars (Cropper, 1990; Thorstensen et al., 2020), the WD rotates synchronously with the orbital period causing the accretion flow to always interact with the same field lines (see Figure 2.22). Polars get their name from the linearly and circularly polarised light they produce. This polarisation can be used to deduce the geometry of the accretion (Hellier, 2001). The majority of polars exist below the period gap, though generally, they lie within the period range of 80 minutes to 4 hours (Thorstensen et al., 2020).

Photometric properties

As the matter stream impacts the white dwarf surface at $\sim$ 3000 km/s, an accretion column forms that extends to $0.1\,R_{WD}$ above the WD surface. Both soft and hard X-ray emission are generated in this region, with the majority being soft X-rays. These

originate from dense blobs of matter that plunge deeply into the white dwarf's surface. Their kinetic energy is absorbed by the white dwarf and subsequently percolates to the surface, emerging as blackbody radiation at a temperature of 200,000 K (Cropper, 1990). As the accretion spot/column comes in and out of view during the system orbit, the emission may be observed to vary at the orbital period (Hellier, 2001; Thorstensen et al., 2020). The obscuration of the accretion flow or spot behind the limb of the WD and/or donor star can also lead to optical variability on similar timescales (Thorstensen et al., 2020). Polars are also characterised by long-term variations in the total brightness, where they switch between high and low states. AM Her shows a mixture of long and short, low and high states. Low states can last several days to months, while high states may last several months to years with no obvious pattern. This kind of behaviour is evident in many polars with brightness ranges of several magnitudes (Sun et al., 2021; Kalomeni, 2012) (see Figure 2.23). Due to the absence of an accretion disk, the cause of low states in polars has been attributed to a suppression of mass outflow from the donor. Mechanisms put forward to account for this suppression include: a change in the topology of the magnetic field at the L1 point set by the WD and donor (Wu & Kiss, 2008); the migration of star spots underneath the L1 Lagrangian point (Livio & Pringle, 1994); and an interaction between the fields of the donor and WD in the presence of starspots (Duffy et al., 2022).

2.4.4.3 Intermediate Polars

Intermediate polars (or DQ Herculis stars; Patterson 1994; Ramsay et al. 2008) represent the intermediary between polars and non-magnetic CVs with magnetic field strengths of between 1 and 10 MG. The radius of the magnetospheric boundary is believed to be smaller than the circularisation radius, therefore a partial accretion disk may form with the inner disk truncated by magnetically controlled accretion (see Figure 2.24). The field strengths of the WD are insufficient to cause synchronous rotation, with spin periods typically within the range $0.01\,P_{orb} < P_{spin} < 0.1\,P_{orb}$ (de Martino et al., 2020). Intermediate polars mostly lie above the period gap, with the majority of sources residing at 3–6 hours. Intermediate polars tend to produce harder (more energetic) X-rays than their polar counterparts. This emission is due to accreting material being channelled onto the magnetic polar regions of the white dwarf, where a strong shock develops. The resulting hot post-shock gas cools via thermal bremsstrahlung radiation as it settles onto

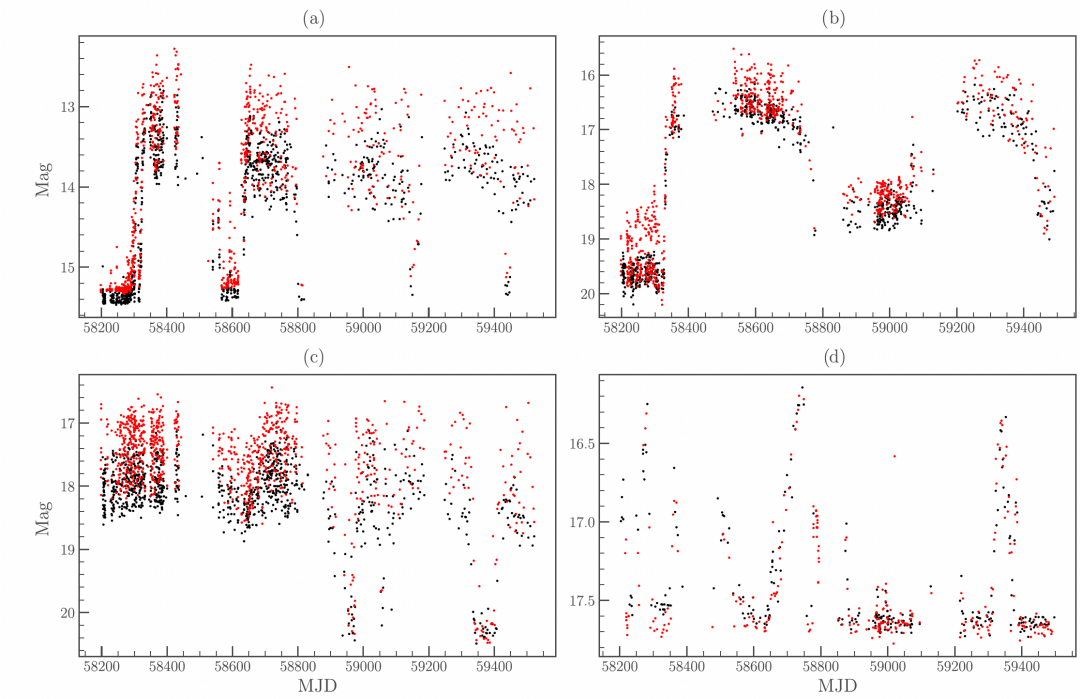


FIGURE 2.23: Light curves of four polar CVs taken with ZTF data coloured red and black to denote r and g band photometry, respectively (Duffy et al., 2022). (a) AM Her with both long and short-duration states, (b) SDSSJ154104 + 360252 shows only long-duration state changes, (c) MT Dra shows only short-duration state changes, and (d) AP CrB shows only short-duration state changes to a higher state.

the white dwarf surface, producing the observed hard X-rays (Patterson, 1994; Anzolin et al., 2008).

Misalignment between the magnetic and spin axes of the WD causes the strength of the magnetic field at a given radius to vary with the spin cycle. The material flowing to the upper pole is picked up from the region of the disk to which it points, while the material flowing to the lower pole is picked up from the opposite side of the disk (see figure 2.25). Accretion onto both poles creates a fundamental difference between intermediate polars and polars. For a two-pole accretor, when accretion onto one pole is obscured, accretion onto the other will be visible such that the X-ray flux never reaches zero.

Photometric variability

Light curves may contain multiple short timescale periodicities due to the orbital period, spin period of the WD and the beat period between the spin and orbital period. A major contributing factor stems from disk material being fed onto multiple field lines covering a range of azimuth angles. Our changing view of the curtains of matter lifted out of the orbital plane, can produce modulations from X-ray to the optical (Rosen et al., 1988).

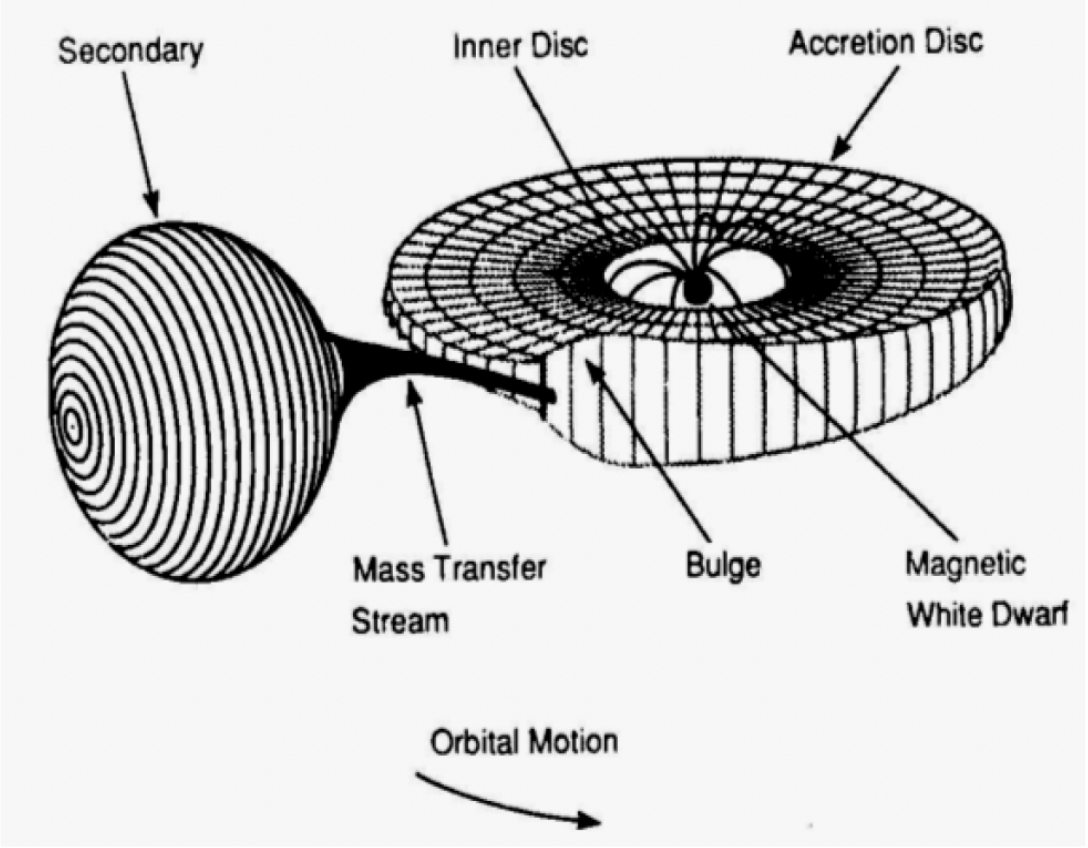


FIGURE 2.24: Schematic diagram of an intermediate polar (Giovannelli, 2008)

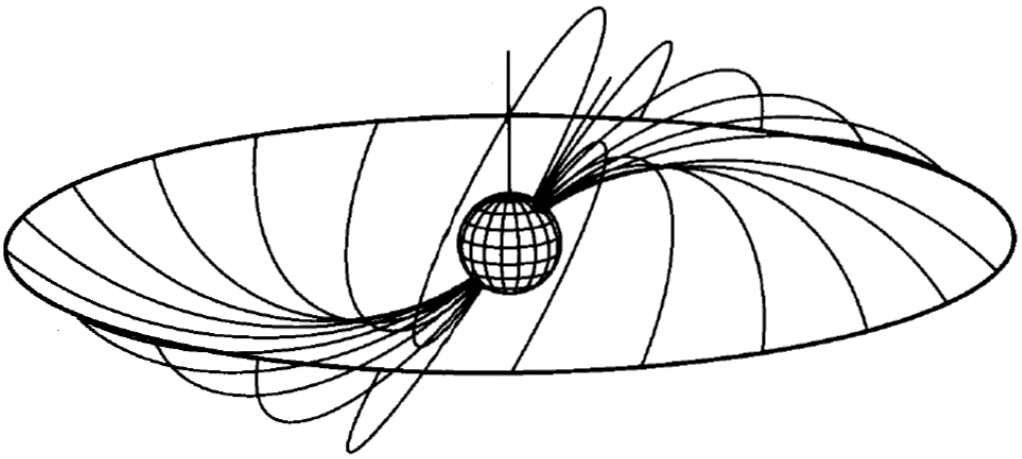


FIGURE 2.25: The pattern of field lines leading from the inner edge of the disk to the white dwarf (Hellier, 2001).

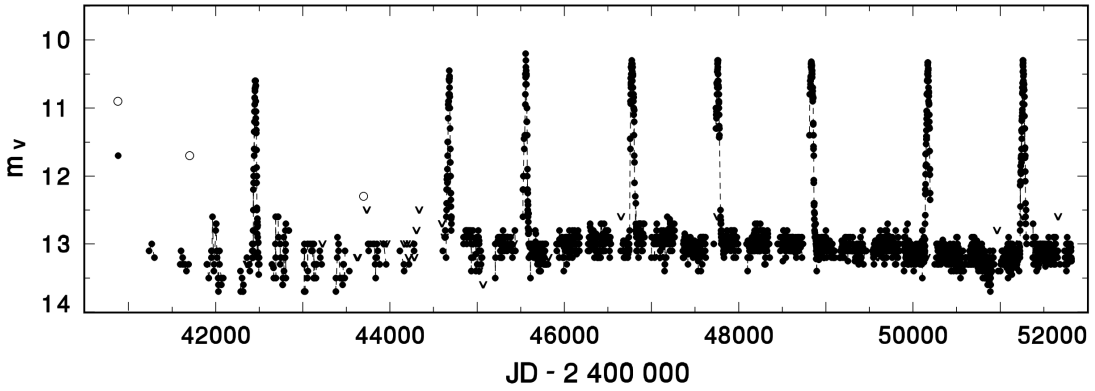


FIGURE 2.26: Optical light curve of GK Per over the years 1970–2000. Upper limits of brightness are represented by the v symbols; empty circles mark the maxima of three outbursts that fall in data gaps (see Šimon 2002 for details).

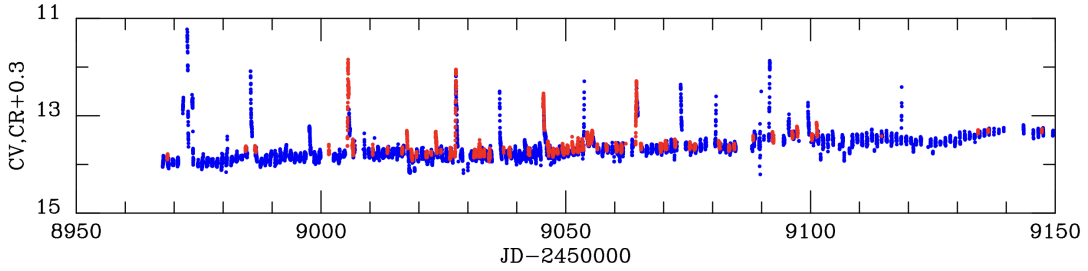
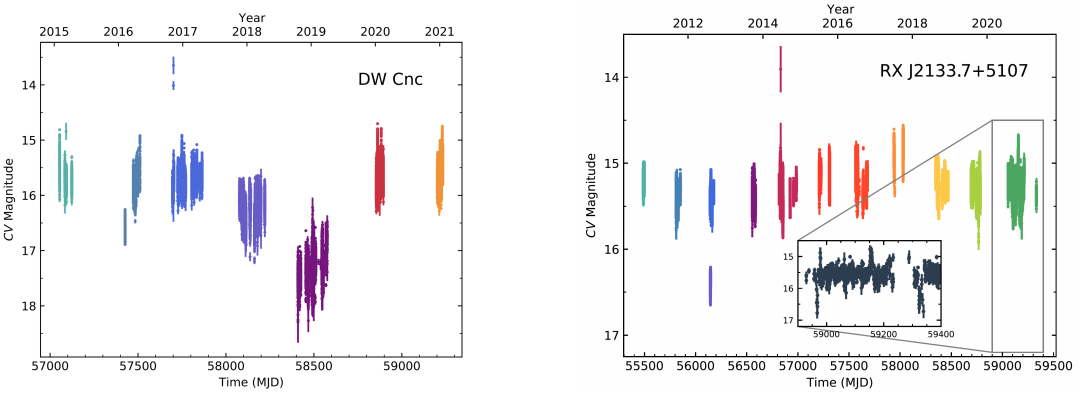


FIGURE 2.27: Light curve of V1223 Sgr (data from AAVSO). The unfiltered visual magnitude with the Bessel V zeropoint, CV, is shown in blue; the red points show the unfiltered red magnitude CR with Bessel R zeropoint plus 0.3 (Hameury et al., 2022).

According to the application of the DIM to intermediate polars, dwarf nova outbursts should still be possible despite the truncation of the inner accretion disk (Hameury & Lasota, 2017). GK Per, for example, has been observed to undergo 'normal' outbursts recurring every $\sim$ 3 years, with an amplitude of 2–3 magnitudes and durations of 50–60 days (Figure 2.26) (Šimon, 2002). In general, dwarf nova outbursts are not a common feature among IPs (Hameury & Lasota, 2017), however, some IPs display short outbursts (several hours) that cannot be explained by the DIM. For example, for $\sim$ 6 months in 2020, V1223 Sgr underwent a series of outbursts each with a typical duration of several hours and a $\sim$ 6 day recurrence period (Figure 2.27), which Hameury et al. (2022) attributed to the magnetic–gating instability model proposed by Spruit & Ronald (1993). The model describes a repeating cycle in which material accumulates at the disk's inner edge, where the centrifugal barrier created by the WD magnetic field prevents accretion onto the star. As the surface density increases, the material overcomes the barrier and accretes onto the star.



(A) DW Cnc (B) RX J2133.7+5107

FIGURE 2.28: (A) AAVSO light curve of DW Cnc from 2015 to 2021 taken in a clear filter mapped onto the V band (CV). Decline to low-state began around MJD $\sim$ 58080, before reaching its lowest flux (CV $\sim$ 17.5) around MJD $\sim$ 58400. Rise began soon after, recovering to its typical flux (CV $\sim$ 15.5) by MJD $\sim$ 58850. (B) AAVSO light curve of RX J2133.7+5107 from 2010 to 2021. The inset is an ASAS-SN band light curve from 2020 to 2021 with two short-lived drops in flux. The colours represent the different epochs used for timing analysis. From Covington et al. (2022)

.

A more prevalent feature of intermediate polars is the presence of transitions from an average (or high) brightness state to a low state that may last weeks to years with depths of 0.5 magnitudes or more. Such state transitions are less common than in polars (Covington et al., 2022; Šimon, 2021). As with polars, the temporary reduction/cessation of mass transfer due to star spot migration is the most popular theory (Livio & Pringle, 1994). Observed changes in the X-ray and optical light curve periodicities during low states have led to proposals that systems may switch from the typical disk-fed accretion to either purely stream-fed, where the accretion stream flows directly onto the WD magnetosphere (Hellier & Beardmore, 2002), or simultaneous stream and disk-fed accretion (Hellier, 1993) onto the WD (Covington et al., 2022). Figure 2.28 shows examples of systems entering low states of different durations.

2.4.5 Novae

Novae are modelled as thermonuclear runaway events within the accreted layer of hydrogen on the WD surface (e.g., Bode & Evans 2008; Munari 2012; Chomiuk et al. 2020; Darnley & Henze 2020), they produce a sudden high amplitude (8–15 magnitudes typically) increase in optical brightness with a long duration decline (weeks to years). Typically, the donor is a late-type main sequence star, though there is a small group

where the donor may be more evolved, e.g., a sub-giant or red giant, and magnetically controlled accretion plays a role (Darnley et al., 2012). Novae are inherently recurrent with recurrence times largely dependent on the WD mass and donor mass transfer rate. Novae may be grouped by recurrence times. Recurrent novae (RNe) have been observed to undergo more than one nova eruption, with recurrence times below 100 years, while classical novae have only been observed to undergo a single eruption with recurrence time extending up to 100,000 years.

2.4.5.1 The nova eruption

As hydrogen-rich material accumulates on the surface of the WD, it instantaneously becomes electron degenerate due to the compression caused by the strong WD gravity. As the envelope thickens, the conditions of temperature and pressure at its base will reach values sufficient to initiate hydrogen burning first via the proton-proton chain and then via the CNO cycle whose energy generation rate is extremely sensitive to temperature $\epsilon_{CNO} \propto T^{18}$. The burning heats the layer, however, it cannot react by expanding because of the decoupling of temperature from pressure under degenerate conditions. Consequently, an increase in temperature leads to an increase in the nuclear energy generation rate, which in turn leads to an exponential increase in the temperature of the envelope in what is referred to as thermonuclear runaway or TNR.

On a timescale of order seconds to a minute, temperatures reach and exceed $\sim 7 \times 10^7$ K, enabling degeneracy to be lifted. The subsequent recoupling of temperature and pressure causes a violent expansion and cooling of the envelope (Starrfield et al., 2016; Jose, 2016). This causes the envelope to become optically thick generating continuum emission initially observable as a short, bright, soft X-ray flash before the peak of the spectral energy distribution shifts from UV to the optical (Kato & Kojiguchi, 2021). The envelope comprises ejecta with escape velocity and a remainder that recedes back to the WD. The peak of optical emission occurs when together they have the largest radius, after which the ejecta continues to expand becoming optically thin and decouples from the remaining photosphere, which recedes back to the WD. The clearing view of the receding photosphere causes the peak of the spectral energy distribution to shift back from the optical to higher energies. If the ejecta become transparent before the shell nuclear burning has ceased, the super-soft X-rays source (SSS) may be revealed (Hachisu et al.,

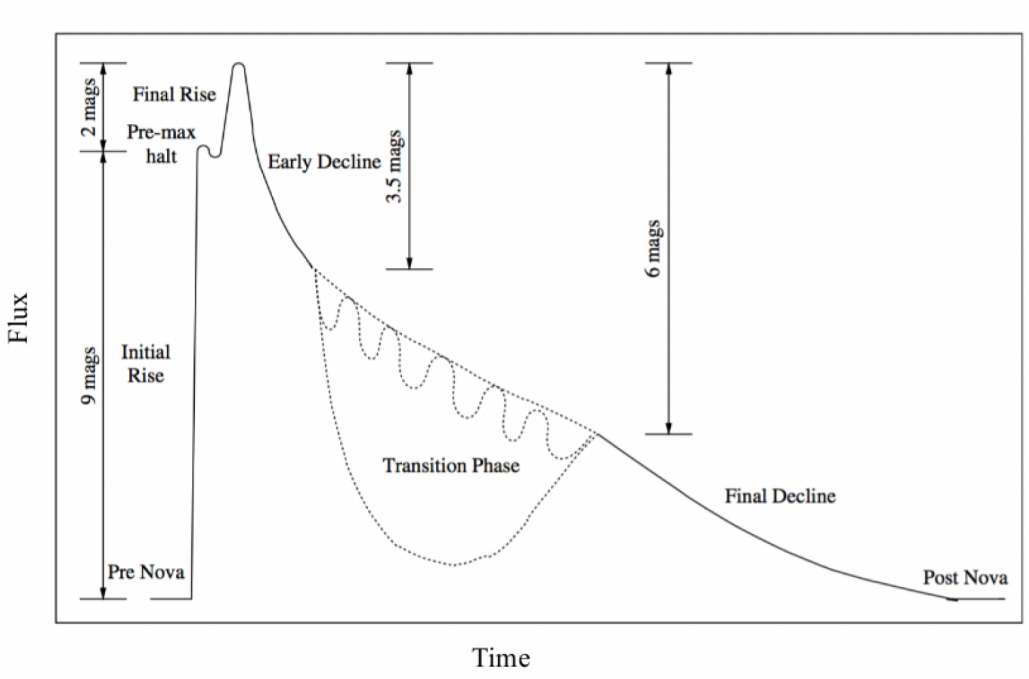


FIGURE 2.29: Morphology of an optical light curve of a typical nova (Bode & Evans, 2008).

2006; Krautter, 2008), a phase that may continue for weeks to decades (Henze et al., 2014; Kato & Hachisu, 2020).

2.4.5.2 Photometric behaviour

A schematic diagram for a nova is shown in Figure 2.29. After an initial rapid rise (a few hours to a day) to peak, nova light curves are well described by a broken power law, where a rapid decline is followed by one that is more gradual. Peak absolute magnitudes can be anything within the range of $-10 < M_V < -5$ (Shafter, 2017; Özdönmez et al., 2018). One much-used measure of a light curve's properties is the time it takes to decline by 2 or 3 magnitudes (t_2 or t_3) from peak with values that can range from $t_2 < 10$ days to $t_2 = 150-200$ days for the very fast and very slow speed classes, respectively (Shafter, 1997; Burlak & Henden, 2008; Payne-Gaposchkin, 1964).

A more detailed look at the light curves reveals an initial rise to within 1-2 magnitudes of the nova's maximum luminosity. This is followed by a pre-maximum halt that may last several hours to several days (Hounsell et al., 2016). The light curve will then quickly rise to its maximum with amplitudes typically in the range of 8–15 magnitudes in the optical. The speed class will determine the duration of maximum light, which can be

between hours and days. Following maximum, fast novae show an early decline, while slow novae may show some oscillations. Some may experience a large dip in the light curve due to dust formation in the ejecta that absorbs optical emission and re-emits in the infrared. Comparison of optical and infrared light curves during this time will show an anti-correlation. A catalogue of 93 well-observed novae (almost all V band) light curves from the AAVSO is presented in Strope et al. (2010), which explains an array of light curve shapes that deviate from a picture of a gradual decline from peak (see Figure 2.30). Of the 93 analysed, 38% followed the smooth declines (e.g. CP Lac, V1668 Cyg, V2275 Cyg) one would expect from the simple light curve model, the remainder showed post-peak features such as:

- Plateus - smooth decline interrupted by a long-lasting nearly flat interval followed by a steeper decline, e.g. V633 Sgr, CP Pup, and RS Orph (21%);
- Dust dips - decline interrupted by sharp dip and recovery to just below the original decline, e.g., DQ Her, FH Ser, V705 Cas (18%);
- Cusp-shaped secondary maxima - secondary maxima with steepening rise then steep decline, e.g., V2362 Cyg, V1493 Aql, V2491 Cyg (1%);
- Quasi-sinusoidal oscillations superimposed on otherwise smooth decline, e.g., V603, GK Per, GK Per, V1494 Aql (4%);
- Flat-topped LCs - smooth light curve with an extended interval at the peak with near constant brightness, e.g., DO Aql, V849 Oph, BT Mon (2%);
- Jitters/flares superposed on the decline — substantial short duration variability brightenings, e.g., DK Lac, HR Del, V723 Cas (16%). Apart from dust-dips, most of the other features lack strong theoretical explanations.

2.5 Spectroscopic properties of CVs

The optical spectra of dwarf nova systems in quiescence will exhibit strong/broad Balmer emission lines originating from an optically thin accretion disk. Less prominent lines of neutral and singly ionised helium may also be present along with further quiescent

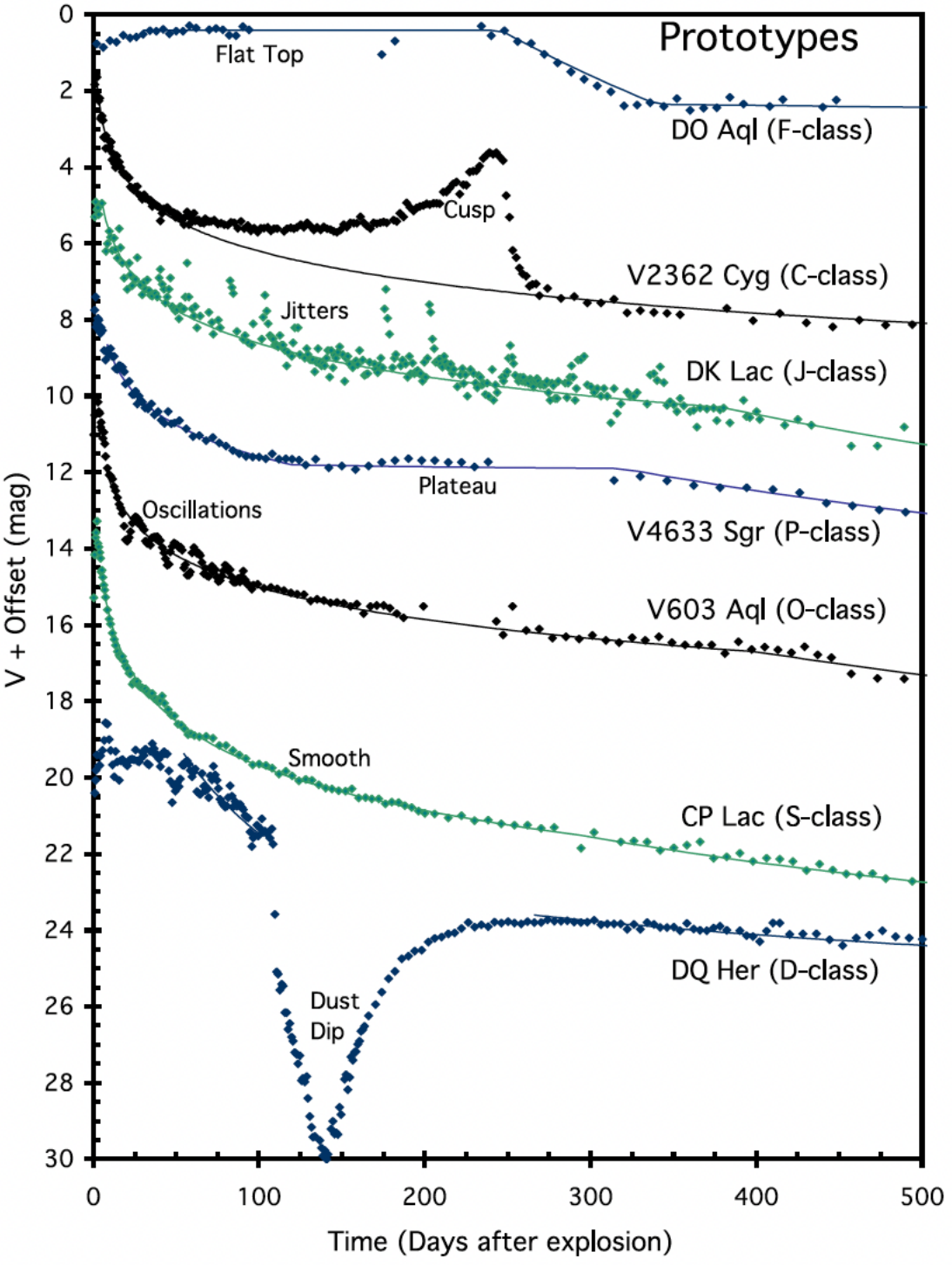
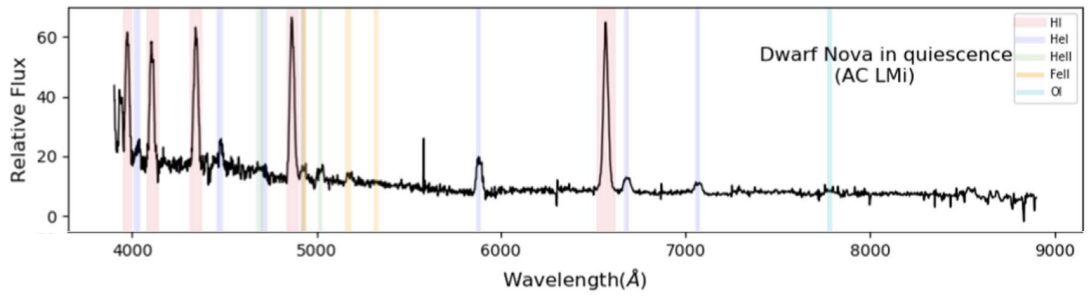
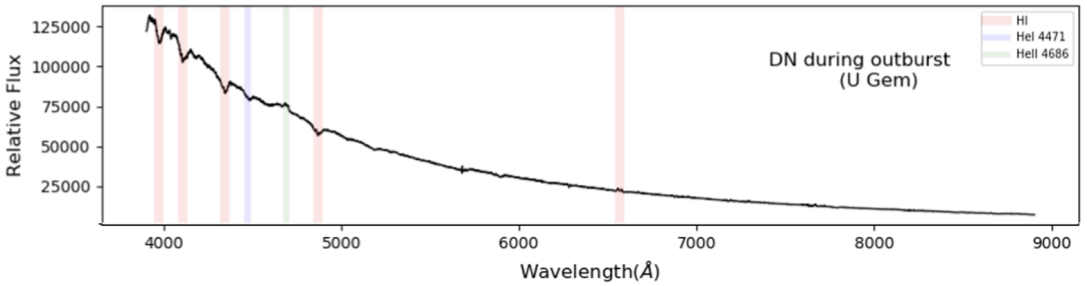


FIGURE 2.30: Examples of nova light curves displaying differences in their post-peak profiles. These differences have allowed nova light curves to be grouped into different categories (see Strope et al. 2010).



(A) AC LMi quiescent spectrum



(B) U Gem outburst spectrum

FIGURE 2.31: Dwarf novae spectra during quiescence and outburst. Spectral lines are marked by different colours for each element, these include HI, HeI, HeII, FeII and OI.

characteristics (Hou et al., 2020). These include: Hβ emission at least ~twice as strong as HeII λ4686; weak or blended CIII/NIII λ4650; and the presence of Fe emission lines such as FeII λ5169, λ5317, and λ4924 blending with HeI λ4922. During outburst, absorption lines of Balmer lines and HeI λ4471 with similar widths as in quiescence, as well as He II λ4686 in emission is present. Narrow emission cores can appear within the broad absorption lines which may indicate a decline from an outburst or a CV system with low disk contribution with emission from the underlying stars producing the luminosity. Figure 2.31 provides an example of a dwarf nova in quiescence and outburst.

Higher excitation conditions are present in nova-likes than in dwarf novae such that HeII λ4686 and CIII/NIII λ4650 features are relatively stronger, and the HeII λ4686/Hβ emission ratio may exceed unity (Warner, 1995; Hou et al., 2020). Apart from FeII emission which is rarely seen, many of the optical lines displayed by dwarf novae are also present for nova-likes, with a few additional weak lines of OII, SiII, CII, and CIV. Some nova-likes may display absorption line spectra, these are usually the UX UMa subtype.

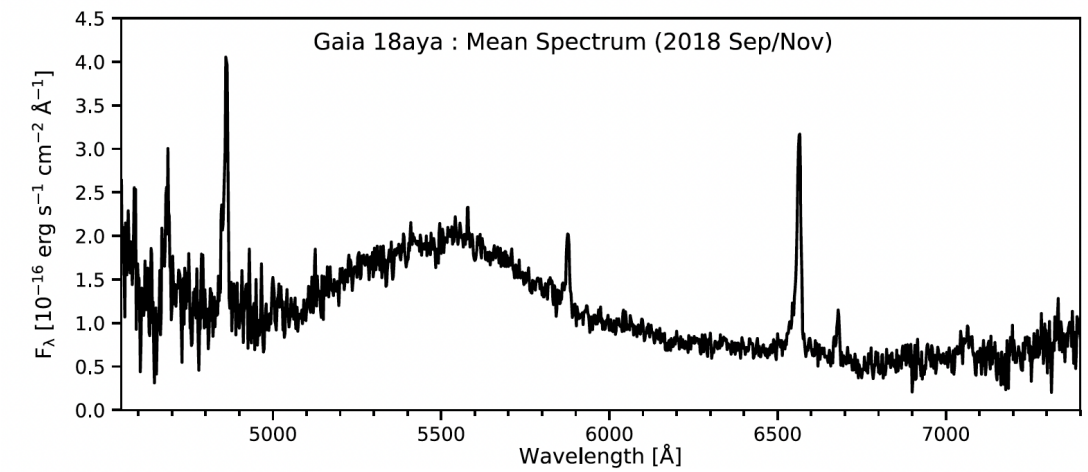


FIGURE 2.32: Mean spectrum of Gaia18aya from 2018 September and November, with the prominent cyclotron hump at $\sim$ 5500 Å. The characteristic Balmer and HeI and HeII emission lines are present.

The optical spectra of AM CVn stars tend to be dominated by helium emission during quiescence and helium absorption during outburst, showing little trace of hydrogen (Solheim, 2010).

The spectra of polars show strong Balmer, HeI and HeII emission along with strong HeII λ4686A emission comparable to Hβ (Thorstensen et al., 2020). The motion of particles in the accretion stream around the field lines produces cyclotron emission at the cyclotron frequency and harmonics thereof. Variations in the strength of the magnetic field with position and time cause cyclotron frequency variation about some fundamental frequency. The cyclotron emission therefore varies about this fundamental frequency and harmonics thereof. This may be seen in the spectra of some polars as cyclotron humps, the presence of which may be used to deduce the strength of the magnetic field (Hellier, 2001) (see Figure 2.32).

The spectral lines of intermediate polars generally resemble polars (Warner, 1995; Hou et al., 2020). Aside from the usual Balmer emission lines, prominent lines of HeI, HeII, and the CIII/NIII λ4650 blend will be present. The strength ratio Hβ/HeII λ4686 can be used to separate intermediate polars from polars. For polars, these two lines are comparable in strength, while for intermediate polars, HeII λ4686 is generally slightly weaker than Hβ.

All novae show Balmer lines. There are two distinct types of nova spectra (seen during the rise and early decline): FeII spectra show numerous singularly ionised iron lines

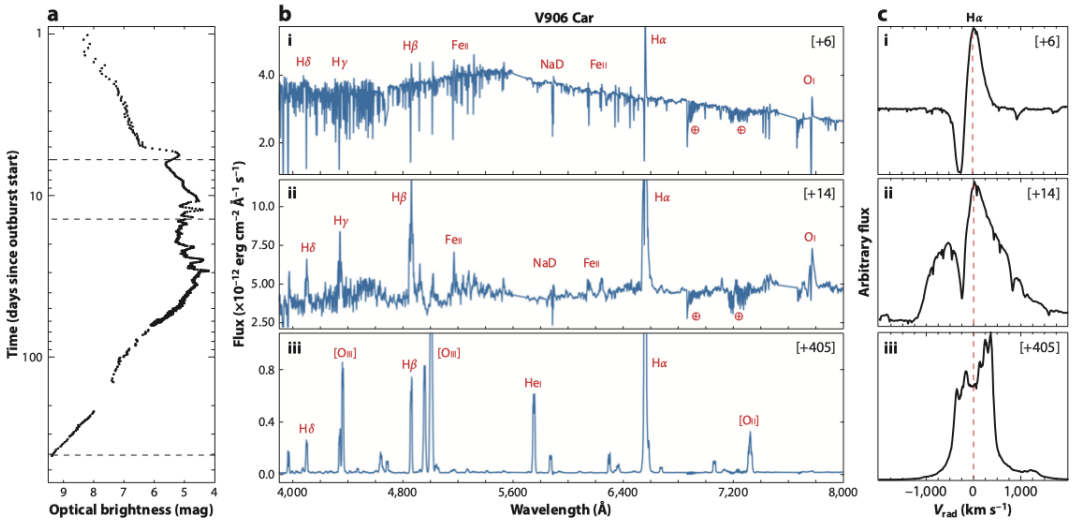


FIGURE 2.33: The optical evolution of nova V906 Car (a) the optical (R-band) light curve, (b) optical spectra, and (c) close-ups detailing spectra around the Hα line. Panel (b) is split into three subpanels, i, ii, and iii, marking 5 days before light curve maximum (peak), 3 days after peak, and more than a year after maximum. (b,i) shows a photospheric spectrum with relatively narrow P Cygni profiles.(b,ii) shows the strengthening and broadening of emission lines, with absorption components from the previous spectrum still superimposed on the emission lines. (b,iii) shows a nebular spectrum dominated by high-excitation and forbidden emission lines. Panel (c) is split similarly (Chomiuk et al., 2020).

with velocities < 2500 km/s, believed to have formed because the ejecta have ploughed into the secondary/circumbinary material; He/N spectra display HeI/II and NIII lines with velocities > 2500 km/s which occurs due to unimpeded ejecta. Nova spectra make the transition from being absorption line-dominated to emission line-dominated as the ejecta make the transition from being optically thick to optically thin. P Cygni profiles are present as the system rises to optical maximum (Chomiuk et al., 2020). An example of the optical evolution of FeII type nova V906 Car is provided in Figure 2.33.

2.6 Examples of active research areas

2.6.1 Disk Instability Model

The research into CVs has led to a significant advancement in our understanding of binary evolution and mass transfer, and has opened the door to many interesting areas of research. For example, the DIM, and variants thereof, help explain dwarf nova outburst diversity. Adjustments to the model even provide a possible explanation for outbursts in

a subclass of low-mass X-ray binaries — soft X-ray transients — that consist of a neutron star accreting from a low-mass main sequence companion (Hameury, 2020). However, despite its successes, a satisfactory variant of the DIM has yet to be found to explain the outburst profiles of WZ Sge systems (Kato, 2015) or the diverse outburst behaviour seen in AM CVns (e.g., Rivera Sandoval et al. 2022; Duffy et al. 2021). In the latter case, Duffy et al. (2021) studied systems with $22.5 \leq P_{\rm orb} \leq 26.8$ minutes and found that AM CVns deviated from the expectation that systems with similar orbital periods exhibit similar outburst activity. They proposed that the uncertain nature of the donor star or the formation channel is a major contributing factor (another important research avenue).

2.6.2 Mass Growth in White Dwarfs

On the topic of CV evolution, specifically type Ia supernova progenitors, the validity of the single degenerate pathway hinges on whether the white dwarf will eject less mass than is accreted at the end of each nova cycle (time between successive nova eruptions), and subsequently grow to the Chandrasekhar mass limit ($1.4\,{\rm M}_{\odot}$). Many such studies have been undertaken to ascertain the possibility of WD mass growth with recurrent novae seen as strong candidate type Ia progenitors, these are described as possessing high mass WDs accreting at high rates. Work by Kato et al. (2015) supports the possibility of one such system, M31N 2008-12a (or '12a'; Darnley et al. 2016, 2015) as a Ia progenitor. Kato et al. (2015) calculated the supersoft X-ray source (SSS) phase duration for several white dwarf masses under the condition of a 1-year recurrence period (that of '12a'). The SSS phase is the period under which the ejecta from the eruption becomes optically thin allowing us to see the supersoft X-ray emission from the steady hydrogen burning on the WD surface. For the 2014 eruption of M31N 2008-12a, the SSS phase lasted 12 seconds. For this duration, the white dwarf mass would need to be $1.38\,{\rm M}_{\odot}$. After modelling the optical/UV and supersoft X-ray light curves based on this mass and an accretion rate of $1.6 \times 10^{-7}\,{\rm M}_{\odot}{\rm yr}^{-1}$, the ejected mass was calculated to be $6 \times 10^{-8}\,{\rm M}_{\odot}$. Thus Kato et al. (2015) found the white dwarf is increasing in mass over each nova cycle. Hillman et al. (2015) conducted numerical hydrodynamical simulations to identify the range of accretion rates under which the WD mass may grow towards the Chandrasekhar limit. They found that accretion rates in the range $0.3 - 6.0 \times 10^{-7}\,{\rm M}_{\odot}{\rm yr}^{-1}$ led to gradual mass growth for a $1.4\,{\rm M}_{\odot}$ WD, while WDs in the range of $1.0 - 1.4\,{\rm M}_{\odot}$ accreting at

$5 \times 10^{-7}\,M_{\odot}yr^{-1}$ were also found to grow in mass. In an extension of this research Hillman et al. (2016) incorporated helium shell ejections that arise due to the products of previous shell hydrogen burning episodes from past novae. They found mass growth is possible despite such occurrences. However, as Kato et al. (2015) pointed out, the gradual reduction in accretion rate as the system evolves is not accounted for in these simulations.

2.6.3 Transitions Between High and Low-States

The physical origin of the transitions between high and low states of brightness with no discernable pattern is poorly understood, particularly for magnetic CVs. Livio & Pringle (1994) suggested the cause of state transitions to be the migration of star spots to regions close to the L1 Lagrangian point causing a reduction in the mass-transfer rate. In an extension of this theory, Wu & Kiss (2008) proposed that low states in AM Her, the archetypal polar, were due to a realignment of the system's magnetic field in response to the change in mass-transfer rate. However, it does not explain how low-mass M stars generate sufficient amounts of star spots, and what would cause their migration towards the L1 region (a topic explored in Hessman et al. 2000). Duffy et al. (2022) examined polar photometry of several systems finding that short-lived states are a relatively common occurrence and proposed that they are due to an interaction between the magnetic field of the white dwarf and donor star spots.

2.6.4 Expanding the Cataclysmic Variable Sample

The above serves to highlight the importance of this class of transient, in fact, the discussed research areas represent only a fraction of the numerous open questions surrounding these systems. If we are to constrain models of binary evolution and accretion, and better understand the physical origins of the observable characteristics of CVs, one requires a large sample of such objects, especially rare types such as AM CVns and strongly magnetic systems. Since properties such as orbital period, mass-transfer rate, accretion rate, donor composition, and orbital inclination form a continuum of values, a greater CV sample size is important to explore the true diversity of examples associated with each class of CV and also those examples that exist at class boundaries in such a parameter space.

Chapter 3

Machine Learning

3.1 Introduction

Machine learning (ML) is a field of Artificial Intelligence focused on algorithms that learn patterns from data to make predictions (Hastie et al., 2003). Several fields of ML exist, with the fields of supervised and unsupervised learning being the main focus of this thesis. Supervised learning involves learning a mapping between a set of input variables, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_N]$, and output variables, $\mathbf{Y} = [y_1, y_2, y_3, ..., y_N]$, such that given new input data, $\mathbf{X}'$, prediction of the output variables $\mathbf{Y}'$ can be made. Supervised learning can be split into the tasks of classification and regression. Classification refers to problems where the output variable, $\mathbf{y}_i$, corresponding to the example, $\mathbf{x}_1$, is one of a set of class labels (e.g., dog, cat, squirrel), whereas the output is a non-discrete (continuous) variable in the case of regression. Unsupervised learning focuses on identifying inherent patterns, relationships, or structures within the data without using output labels. The branch of unsupervised learning I focus on in this work is dimensionality reduction. Dimensionality reduction is concerned with projecting data that exists in a high-dimensional space onto a lower-dimensional plane (usually 2D or 3D), such that examples that are close together in high-dimensional space are also close in the lower-dimensional space, thereby conserving the relationship between examples. It is useful for viewing relationships between examples in a lower-dimensional space.

3.2 Input Data

Machine learning algorithms can handle various types of data, each with its own set of characteristics and challenges. Here, I briefly discuss three common types of data used as input in machine learning and the role of feature engineering in this context.

3.2.1 Data Types

Image data, presented as a grid of pixels, is prevalent in the field of computer vision, where algorithms such as convolutional neural networks (CNNs; Kwak 2016) process pixel values and their spatial relationships to extract meaningful information such as patterns and objects within images. This extracted information can then be used for object detection, image classification, or scene understanding (Chollet, 2021). Time series data consists of observations collected at regular or irregular time intervals (e.g., weather data, astronomical light curves). Algorithms such as recurrent neural networks (RNNs; Sherstinsky 2020) and variants thereof are effective at capturing patterns and trends by processing such data sequentially and retaining information from previous observations. Structured data tables are widely used in ML. They are organised such that rows represent data instances, while columns represent attributes (or features) of each instance. The generation of features involves the field of feature engineering which includes, but is not limited to, the process of extracting useful information from raw data (e.g., time series data), and manipulating such data to generate features that can be used as input for algorithms.

3.2.2 Astronomical time series data representations

Astronomical time series data of the kind used in this research is irregularly sampled due to factors such as weather conditions, seasonal gaps, instrument availability, observing schedules, and limiting magnitude. Furthermore, lengths of light curves also differ, sometimes by as much as several hundred data points. Several methods have been developed to effectively compare light curves from different astrophysical sources such that these factors do not unduly influence machine learning models.

A widely used method is to extract features from the light curves such that the data is presented to algorithms in tabular form. Richards et al. (2011) contains an extensive set of features that are robust in the presence of the kinds of heterogeneities present in astronomical time series data. They were used in the classification of variable stars. Such work forms the basis for many of the features present in feature extraction packages such as FATS (Feature Analysis for Time Series; Nun et al. 2015) and FEETS (FEature Extractor for Time Series; Cabral et al. 2018). The features comprise those that describe statistical properties, percentile-based features, as well as periodicities within the light curves. Examples include Fourier component extractors that identify amplitudes and phases of frequency components and their harmonics from the Lomb Scargle Periodogram; the ratio of magnitude percentile ranges; colour, where multiband photometry is present; and simple measures of magnitude variability such as kurtosis, skewness and amplitude.

Another, and more recent alternative is 'dmdt' mapping introduced by Mahabal et al. (2017). This is a two-dimensional mapping of the light curves whereby for each pair of points the change in magnitude (dm) and change in time (dt) is calculated. These dmdt pairs are then binned into ranges of dm and dt to generate a 2D histogram. Each 2D bin corresponds to a pixel within a grid, whose pixel intensities are based on counts. These dmdt representations of the data serve as input for CNNs that automatically extract salient patterns in a form of automated feature extraction that can be used for training and prediction.

3.2.3 Train, test, validation sets

The dataset is usually split into separate training, validation and test datasets. The training set is used to train machine learning models using the algorithms. The validation set is used to test the performance of the model. From the performance of the model as tested on the validation set, adjustments are made to the model in the form of algorithm-specific parameters (or hyperparameters) for fine-tuning. The validation set is only used to check performance to make further adjustments to the model. The test set takes no part in the training or model tuning process and is used to assess the generalisation error of the model — how well the model performs on completely unseen data.

The above splitting is typically performed before any pre-processing, feature selection, or data augmentation procedures are performed to prevent the risk of data leakage (Singhi & Liu, 2006; Demircioğlu, 2021). Data leakage occurs when training data contains information that would not be present when the model is used for real-world prediction. This can lead to biased and inaccurate estimations of model performance. By conducting the train-test split before any data manipulation steps, and only performing the fitting procedure for such steps on the training data, we preserve the integrity of the training process.

3.3 Algorithms

Before delving into data preprocessing, feature selection, and data augmentation techniques used before model training, the algorithms used in this research are introduced along with their associated hyperparameters — parameters that control how the algorithms learn from the dataset, set before the learning process. Tuning hyperparameters can improve model performance whilst reducing the risk of overfitting (i.e. learning the noise in the training data).

3.3.1 Decision Tree-based Ensemble Learning

Algorithms such as Random Forest, AdaBoost, and XGBoost that are used in this research are built with an ensemble of Decision Trees (DT; Rokach & Maimon 2008). In the task of classification, given a dataset consisting of features (characteristics) describing each example within the dataset, and an associated classification, the Decision Tree will recursively perform binary partitions of the dataset based on features and associated thresholds in a way that the class homogeneity of resultant subsets (or nodes) is maximised. Gini impurity or entropy guides the selection of the best feature and threshold at each node. Gini impurity measures class impurity using the formula:

$$Giniimpurity = 1 - \sum_{i=1}^{c} p_i^2 \tag{3.1}$$

where p_i represents the proportion of samples belonging to class i, and c is the number of classes. Lower values indicate nodes purer in class — less mixing or diversity of

class labels. Entropy (Equation 3.2) acts similarly, where higher values indicate higher disorder — classes are more evenly distributed.

$$Entropy = -\sum_{i=1}^{c} p_i \log_2(p_i) \tag{3.2}$$

For each feature and all possible thresholds to split the data, the impurity of the subsets of each potential split is calculated. To measure the quality of a split via Genie impurity or entropy, the weighted sum of the impurities/entropies of the child nodes are compared to the impurity/entropy of the parent node to measure the decrease in impurity/entropy. The weights are proportional to the number of samples in each node. Where entropy is used, the subtraction of the entropy of the child nodes from the parent is typically referred to as the information gain. The combination of feature and threshold which brings about the greatest reduction in impurity/entropy is used. The recursive partitioning continues until the stopping criteria is met, such as the minimum number of samples in a node, or no further decrease in impurity is possible. The resulting tree structure (model) serves to predict class labels of new examples. To do so, new examples traverse the tree based on their feature values, arriving at a leaf node (a terminal node without child nodes). The predicted class label is determined by the mode of class labels within the leaf node of the trained model, while the probability of belonging to that class is computed as the proportion of instances within that node belonging to the predicted class.

The most important hyperparameters of DTs are: the maximum depth of the tree (*max_depth*), where the depth is the number of decision nodes from the root node to the farthest leaf; the minimum number of samples required to split a node (*min_samples_split*; the minimum number of samples required to be a leaf node (*min_samples_leaf*); and the maximum number of features to consider when making a split (*max_features*). Reducing the *max_depth* increases computational efficiency while also reducing overfitting. *min_samples_split* and *min_samples_leaf* also help control overfitting. Adjusting *max_features* helps to increase the diversity of the trees.

Random Forest

Random Forest (RF; Breiman 2001) operates by employing a voting mechanism, using predictions generated by multiple uncorrelated Decision Trees. The class with the

highest number of votes becomes the prediction of our model. Using the bootstrap aggregation technique, each tree in the ensemble is trained on a sample drawn with replacement (i.e., a bootstrap sample) of the original training set. Additionally, a random subset of features is used during this process to ensure the trees remain uncorrelated. Several crucial hyperparameters come into play, these are those associated with DTs with a notable addition - the number of trees in the ensemble. Increasing the number of trees enhances the model's ability to generalise to new data, albeit at the expense of added complexity and computational time.

Adaboost

AdaBoost (ADB; Freund & Schapire 1997) combines Decision Trees sequentially. It is designed to improve the performance of each successive tree by iteratively focusing on instances that are difficult to classify using a weighting mechanism. The procedure starts by assigning uniform weights, w_i to each example in the original dataset such that $w_i = 1/N$, N being the number of examples. A bootstrapped sample with weighted sampling is then generated. On the first iteration, the weights are all equal, therefore, each sample has an equal chance of being selected. AdaBoost trains a DT model and then evaluates its performance by calculating the weighted error — the sum of the weights of the misclassified samples. Based on its performance, an importance is assigned to the model; a model with a lower weighted error is given a higher importance in the ensemble based on the formula:

$$LearnerImportance = \frac{1}{2} \log \left(\frac{1 - weightederror}{weightederror} \right)$$

Next sample weights are updated such that the weights of incorrectly classified samples are increased while those of correctly classified examples are decreased. Such adjustments ensure that subsequent trees focus more on previously misclassified examples. The process is then repeated for as many iterations (or boosting rounds) as specified. The final prediction is computed as the weighted sum of the predictions of all the trees.

Hyperparameters for AdaBoost are the same as for DTs with the addition of *n_estimators* that specifies the number of iterations to perform the boosting over; and *learning_rate* applies a weight to each tree at each boosting iteration, where a higher learning rate increases the contribution of each tree. Increasing *n_estimators* can improve performance

though at the expense of an increase in computation time and may lead to overfitting. Lower *learning_rate* require more trees (iterations) but may improve the generalisation ability of the model.

Extreme Gradient Boosting (XGBoost)

XGBoost (Chen & Guestrin, 2016) is another example of sequentially combining Decision Trees (or weak learners). Where ADB uses weights to improve performance, XGBoost employs a gradient-boosting approach. In gradient boosting, each new weak learner is trained to minimise some loss function (that describes the classification performance) with respect to the previous ensemble's predictions. In this way, XGBoost iteratively improves its performance with each tree addition by utilising information from the prior round's prediction accuracy. The final prediction model is the sum of M weak learners, $F(x) = \sum_{i=1}^{M} f_i(x)$, where $f_i(x)$ corresponds to weak learner i trained on data x. XGBoost utilises parallelised tree building and hardware optimisation to improve runtime, and regularisation to reduce overfitting. Delving deeper, the algorithm follows these basic steps:

Step 0) Given a training set of N examples, $[x_1, ..., x_N]$, and corresponding labels $[y_1, ..., y_N]$, a differentiable loss function $L_i(y_i, \hat{y}_i)$, where $\hat{y}_i$ is the model prediction for training example x_i, and M weak learners, the algorithm is initialised by a base model, $f_1(x)$, such that the predictions are the same value of class probability for each example.

Step 1) Compute the derivative of the loss, L, of $f_1(x)$ for each instance, $r_{i,j}$, these are referred to as pseudo-residuals, where j represents the iteration step or model number.

$$r_{i,j} = -\frac{\partial L_i(y_i, \hat{y}_i)}{\partial \hat{y}_i}$$

Step 2) Train a weak learner (DT) on a dataset where the target values are replaced by the pseudo-residuals $\{(x_i, r_{i,j})\}_{i=1}^{N}$ of the previous model. This produces our next model $f_2(x)$.

Step 3) We can then add some contribution $\hat{\gamma}_2$ of $f_2(x)$ to $f_1(x)$ to produce our new ensemble $F(x) = f_1(x) + \hat{\gamma}_2 f_2(x)$. The contribution is determined using:

$$\hat{\gamma}_2 = argmin_\gamma \left[\sum_{i=1}^{N} L\left(y_i, f_1(x_i) + \gamma f_2(x_i)\right) \right]$$

Steps 1-3 are repeated for M iterations such that $F(x) = f_1(x) + \hat{\gamma}_2 f_2(x) +, ..., + \hat{\gamma}_M f_M(x)$.

Hyperparameters include those mentioned for RF with the addition of parameters such as the learning rate that controls the loss function step size at each iteration, and the regularisation rate to adjust model generalisation — regularisation adds a penalty term to the loss function proportional to the absolute values or squared values of the model parameters.

Feature Importance

Decision Tree methods are very useful for identifying the most relevant features for classification. Feature importance scores may be obtained from the trained model. The scores are a measure of how much each feature contributes to decreasing the class impurity at each node in the decision trees. The features leading to the greatest decrease in impurity, when considering all trees in the ensemble, are considered more important.

3.3.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA; Hastie et al. 2003) is a dimensionality reduction technique also used for classification purposes. Class predictions are obtained using Bayes' rule by finding the class, k, that maximises the posterior probability:

$$P(y = k|x) = \frac{P(x|y = k)P(y = k)}{P(x)} = \frac{P(x|y = k)P(y = k)}{\sum_l P(x|y = l) \cdot P(y = l)}$$

Class distributions are modelled as multi-variate Gaussians assumed to have the same covariance matrix $\Sigma_k = \Sigma$ for each class. This assumption reduces the log of the posterior probabilities to linear functions, which leads to a further assumption, linear separability, since locations where the functions are equal define linear class decision boundaries. This leads to the formula:

$$\log P(y = k|x) = -\frac{1}{2}(x - \mu_k)^t \Sigma^{-1}(x - \mu_k) + \log P(y = k) + constant$$

with sample x and mean μ_k.

3.3.3 Support Vector Machines

Support Vector Machines (SVM; Cortes & Vapnik 1995) works by finding the ideal hyperplane that best distinguishes between two classes in feature space while maximising the margin between the classes under the assumption of linear separability (Figure 3.1). The margin is defined as the distance between the hyperplane and the nearest data points from each class (also known as support vectors). Each data point is represented by a feature vector, $\mathbf{x}_i$, in feature space together with its class label, y_i, whose values are either 1 or -1 to indicate class. To find the ideal hyperplane the optimal weight vector, $\mathbf{w}$, and bias term, b, are found such that $\mathbf{w}^T\mathbf{x}+b = 0$ defines the decision boundary between the two classes. Mathematically, the margin is proportional to $1/||\mathbf{w}||$, so maximising the margin is equivalent to minimising $||\mathbf{w}||$, which is the objective of the classifier. The minimisation can be performed using Lagrangian multiplier methods or gradient descent. Multiclass classification is achieved by splitting data into one class versus all others, performing this for all classes, or framing the problem as multiple cases of one class versus another.

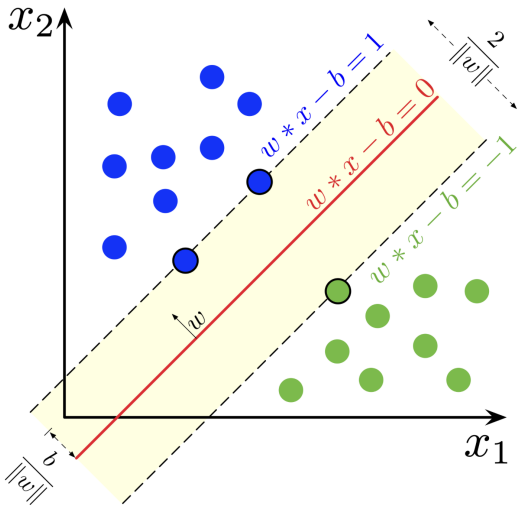


FIGURE 3.1: SVM representation in 2D. The hyperplane is the red line, while margins are represented by a line on either side. Samples on the margin are called the support vectors.

SVM uses the *kernel trick*, to handle non-linear decision boundaries by effectively transforming the data/feature space into a higher dimensional space allowing for linear separation. Mapping to higher dimensional space is computationally expensive, so rather than computing this mapping a kernel function is applied to perform this efficiently and implicitly. The kernel function, $K(\mathbf{x}_i, \mathbf{x}_j)$, computes the inner product of the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the implicit higher-dimensional feature space. A widely used kernel is the Radial Basis Function (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}\right)$.

The most influential hyperparameters are: the choice of kernel, (typically RBF) to model nonlinear decision boundaries; Kernel Coefficient (γ), which governs the influence of individual training examples on the decision boundary, higher γ values produce more complex boundaries, potentially leading to overfitting, while lower values allow better generalisation; and Error Penalty (C), that controls the cost of miss-classification on the training data, a smaller C value yields a softer margin, allowing for more misclassifications but better generalisation, whereas a larger C enforces a hard margin, which may lead to overfitting.

3.3.4 Gaussian Naive Bayes

Gaussian Naive Bayes (GNB; Zhang 2004) is a probabilistic machine learning model used for classification, based on Bayes theorem (Equation 3.3). Given two events A and B, the probability of event A occurring given that event B has occurred (conditional probability) is given by:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{3.3}$$

Where $P(A)$ is the prior probability of event A, which represents our initial belief in the likelihood of A occurring. $P(B|A)$ is the likelihood of observing event B given that event A has occurred. $P(B)$ is the marginal probability of observing event B, which represents the total probability of observing B regardless of the occurrence of A. $P(A|B)$ is the posterior probability of event A given that event B has occurred. It represents our updated belief in the likelihood of A occurring after observing B.

For a machine learning classification dataset of n examples we have $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$ and target labels $\mathbf{y} = [y_1, y_2, ..., y_n]$. The aim is to predict the target label, y, given m features for each data point, $\mathbf{x} = [x_1, x_2, ...x_m]$. The formula can be written as:

$$P(y|x_1, x_2, ..., x_m) = \frac{\prod_{i=1}^{m} P(x_i|y) \cdot P(y)}{\prod_{i=1}^{m} P(x_i)}$$

Here, we make the naive assumption that the variables/features are independent, hence naive Bayes. Values can be obtained by looking at the dataset and substituting them into the equation. The denominator can be treated as a constant as it does not depend on the class label, remaining constant across different classes, such that:

$$P(y|x_1, x_2, ..., x_m) \propto \prod_{i=1}^{m} P(x_i|y) \cdot P(y)$$

For multiclass classification we need to find the class, y, that maximises the posterior probability, therefore the formula for the predicted class, $\hat{y}$, is:

$$\hat{y} = \arg\max_{y} \left(P(y) \cdot \prod_{i=1}^{m} P(x_i|y) \right)$$

For non-discrete feature values (continuous), we assume the values are sampled from a Gaussian distribution, such that the formula for conditional probability changes to:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2} \right)$$

where μ_y is the mean of feature x_i for class y and σ_y^2 is the variance of feature x_i for class y.

3.3.5 K Nearest Neighbours

K Nearest Neighbours (KNN; Zhang 2016) stores the feature-space position vectors of the training set examples. When making class predictions for new examples, it identifies the mode of the classes among the k nearest neighbours from the training set based on

some distance metric, assigning that mode as the prediction for the new example. The hyperparameters that impart the greatest influence on model performance are the number of nearest neighbours, the distance metric (e.g., Euclidean, Manhattan, Minkowski), and the weighting of individual examples such that close neighbours of a query point have a greater influence than those further away.

3.3.6 Artificial Neural Networks

Artificial Neural Networks (ANN; LeCun et al. 2015) comprise interconnected layers of nodes, commonly referred to as neurons (Figure 3.2). This architecture consists of an input layer that receives feature values, an output layer responsible for generating predictions, such as class probabilities, and one or more hidden layers in between. The hidden layers sequentially transform the initial feature values into predictions by applying non-linear functions to linear combinations of previous inputs. The learning process revolves around minimising a loss function, where adjustments to the model parameters are made through an iterative process known as backpropagation in combination with the gradient descent algorithm until convergence to loss minimum is achieved.

Backpropagation computes the gradient of the loss function, $J(\theta)$, with respect to each parameter, θ, using the chain rule of calculus. The gradient descent algorithm utilises these gradients to iteratively adjust the parameters of the model in the direction opposite to the gradient of the loss function with respect to those parameters until convergence is achieved. Mathematically, it can be represented as $\theta = \theta - \alpha \cdot \nabla J(\theta)$, where α denotes the size of the weight update step (learning rate). The adjustments of the model's weights are performed in reverse order from the output layer to the input layer.

The term *Deep Learning* is used where multiple hidden layers are used (hence deep). A multi-layer perceptron (MLP; Figure 3.3) is one of the simplest Deep Learning models consisting of an input layer, several hidden layers and an output layer.

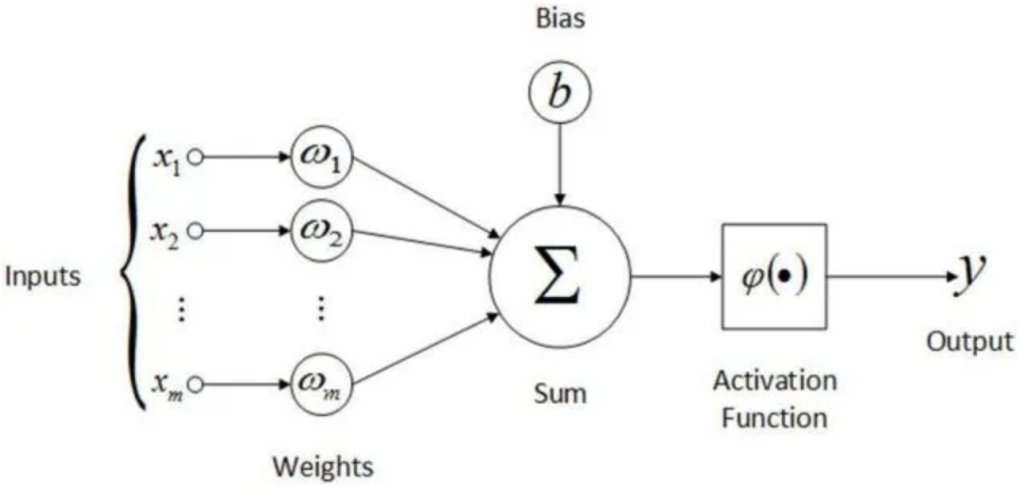


FIGURE 3.2: Neuron architecture. Takes as input a linear (weighted with w_t) combination of the inputs (feature values x_i) along with a bias term (constant b) and puts it through an activation function $\phi(z)$ that introduces non-linearity to produce output y, where $z = x_1x_1, ..., w_mx_m + b$.

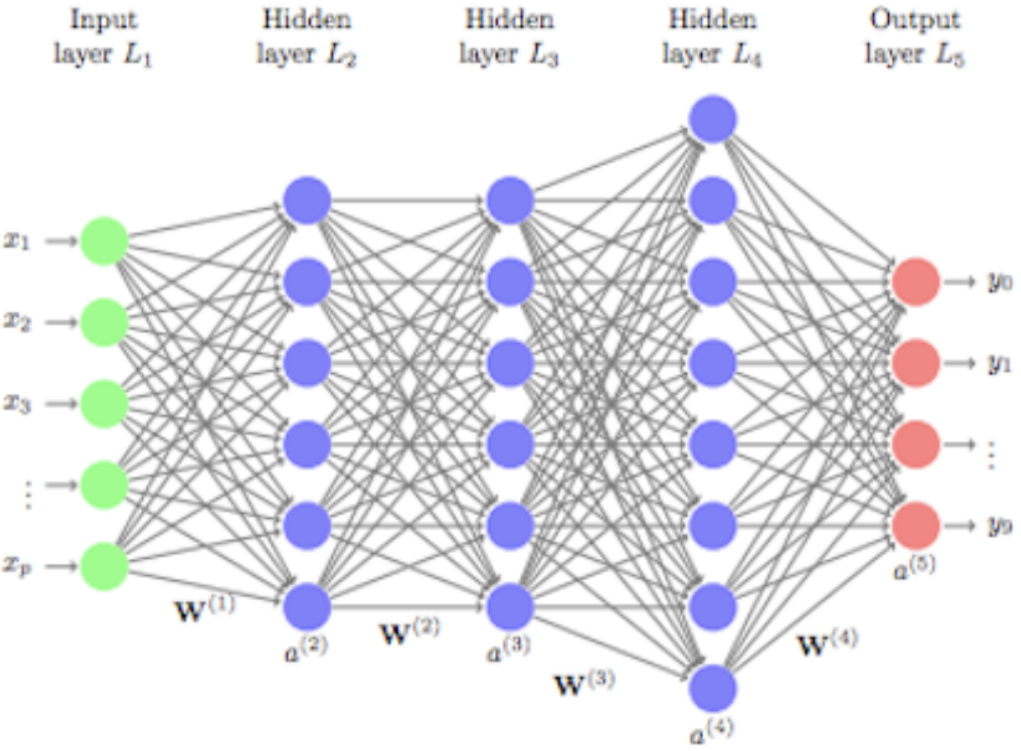


FIGURE 3.3: Architecture of a multi-layer perceptron. Each circle represent a neuron that takes in a linear combination of the previous layers' inputs and passes that through an activation function to produce its output that serves input to each neuron in the following layer (Pérez & Zingaretti, 2019).

To tune the model one usually adjusts the number of hidden layers and the number of neurons per hidden layer. Set these too high (increased complexity) and overfitting may occur. The choice of activation function can impact the capacity of the network to learn complex patterns. Options include the sigmoid function, $\sigma(x) = 1/(1 + e^{-x})$, that outputs values between 0 and 1, and the Rectified Linear Unit (ReLU), $ReLU(x) = max(0, x)$, that outputs 0 for negative inputs and the input value for positive inputs. The learning rate controls the step size of parameters during the gradient descent optimisation process. Higher values will speed convergence to a loss minimum at the risk of overshooting the minimum, while lower values may get caught in a local minimum and slow down training. In addition to these hyperparameters are the batch size — the number of samples to propagate through the network before updating the weights, and the number of epochs — the number of times the entire dataset is passed forward and backward through the network during training. The batch size determined how noisy the updates are with larger values producing more stability but may slow convergence. Training for a larger number of epochs may lead to overfitting.

Convolutional Neural Network

Convolutional Neural Networks (CNN; Kwak 2016) comprise convolutional layers whose outputs feed into an ANN. They are distinguished from regular neural networks by their unique property of translational invariance. This property allows convolutional layers to automatically learn relevant features directly from raw input data, regardless of where those features appear. This ability to identify patterns across different spatial locations eliminates the need for handcrafted features. Input can be image data represented as pixel intensities, image sequences/video, or time series data. Convolutional layers learn local patterns in the data by sliding a grid of elements containing weight values, typically of size 3x3 or 5x5, over the image, stopping at every location. The grid, referred to as a filter or kernel, calculates a weighted sum of image pixel values to produce a feature map (Figure 3.4).

There are multiple kernels for each convolutional layer, the weights associated with each are learnt during the training process such that each resultant feature map captures a different aspect of the data necessary to minimise the error in prediction. Feature maps at each convolutional layer serve as representations of the input data at different levels of abstraction, with initial layers capturing simpler features (e.g., edges, textures)

and deeper layers capturing more complex patterns (e.g., object parts, object presence). Convolutional layers are usually followed by pooling layers that combine elements of the feature maps by averaging or taking the maximum to reduce their size. The output of the final convolution operation (and accompanying pooling) will then be flattened to a vector before input into a standard neural network (Figure 3.5). Hyperparameters of such a network are the same as for ANNs with the addition of the number of filters and their size, and the number of convolutional and pooling layers amongst others.

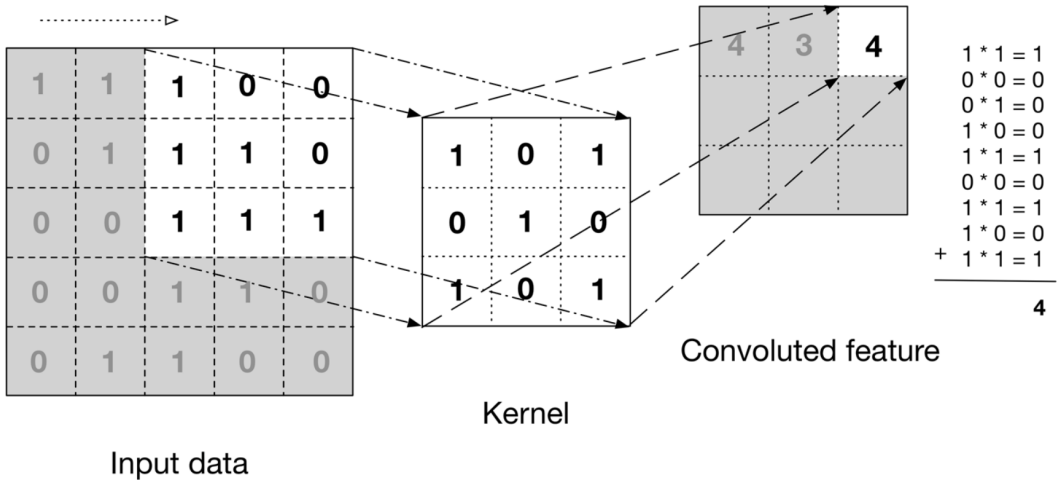


FIGURE 3.4: The convolutional operation uses a sliding kernel, where at each stop element-wise wise multiplication of kernel weights with image pixel values are calculated before summation. The resultant value corresponds to a value in the convoluted feature on the right (Analytics Vidhya, 2021).

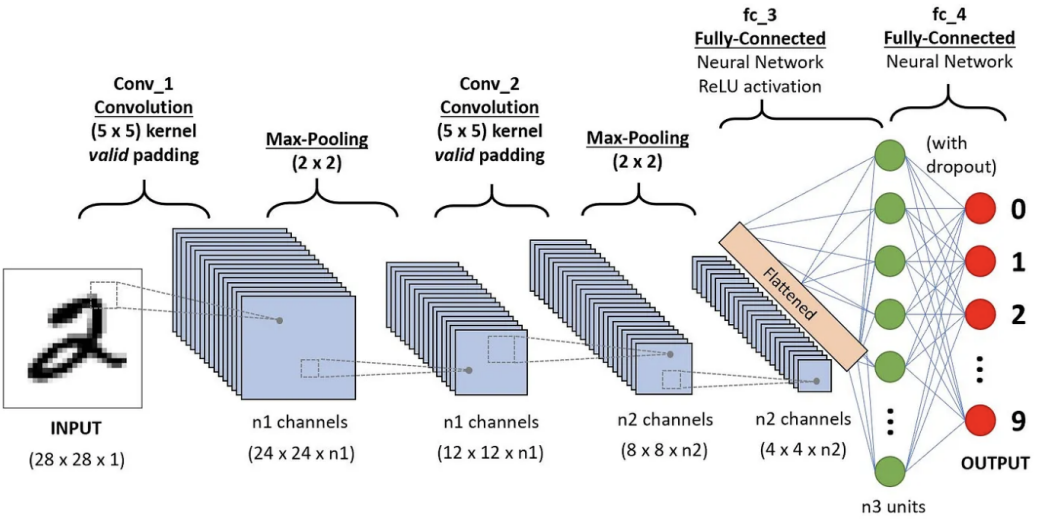


FIGURE 3.5: Example of a CNN architecture with two convolutional layers feeding into a fully connected neural network. Channels refer to the number of feature maps produced which is equal to the number of convolutional kernels (Towards Data Science, 2020).

The application of kernels to local regions of the image, the use of the same weights across different spatial locations of the input, the pooling operation, and hierarchical feature representation due to the use of multiple convolutional layers leads to the translational invariance — the ability to recognise objects or patterns regardless of their exact position in the image. The technique is transferable to time series and video data by changing the dimensions of the convolutional filters to match the input type.

3.3.7 Principal Component Analysis

Principal Component Analysis (or PCA) is a linear dimensionality reduction technique that works by finding the vectors within data space that account for the greatest amount of variance in the data. These will be the eigenvectors of the covariance matrix, Σ, of the feature space, while the magnitudes of the corresponding eigenvalues represent the level of responsibility each eigenvector has in accounting for the variance in the data. Lower dimensional representations of high dimensional data are achieved by plotting examples in the eigenvector space, typically the 2 or 3 which account for the greatest amount of variance.

3.3.8 T-distributed Stochastic Neighbourhood Embedding

T-distributed Stochastic Neighbourhood Embedding (t-SNE; Van der Maaten & Hinton 2008) is a non-linear dimensionality reduction algorithm that takes a high dimensional data space, defined by a dataset consisting of points $X = [x_1, x_2, ..., x_n]$, and transforms it into a lower dimensional representation (or mapping) as output, defined by points $Y = [y_1, y_2, ...y_n]$, aiming to preserve the relationships between data points. This is performed by converting pairwise distances between points in the data space into joint probabilities, p_{ij}, performing similar for a randomly initialised set of matching points in the low dimensional mapping to obtain each q_{ij}, then minimising the Kullback-Leibler (KL) divergence that describes the divergence between the overall joint probability distribution of the data space, P, and that of the low dimensional mapping, Q, through adjustments of the positions of mapping points Y.

The algorithm converts pairwise distances (or similarities) between data space points into conditional probabilities, $p_{i|j}$, such that the similarity of datapoint x_j to datapoint x_i is the conditional probability that x_i would pick x_j as its neighbour if neighbours were chosen proportional to their probability density under a Gaussian centred at x_i:

$$p_{j|i} = \frac{exp[-||x_i - x_j||^2/2\sigma_i^2]}{\sum_{k \neq i} exp[-||x_i - x_k||^2/2\sigma_i^2]}$$

where σ is the variance of the Gaussian centred on x_i. t-SNE symmetrises the conditional probabilities ($p_{j|i}$ and $p_{i|j}$) such that the joint probability p_{ij} reflects the similarity between data points x_i and x_j from both perspectives.

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

A Student t-distribution rather than a Gaussian distribution is used for the lower dimensional mapping to overcome the 'crowding problem' in which moderately separated points in data space become squished together in the low dimensional mapping due to the reduced volume of space available; this is especially problematic for densely population areas of data space. The heavier tail of the t-distribution allows moderate distances in data space to be represented by much larger distances in the mapping compared to

where a Gaussian is used. Using this distribution, the joint probabilities, q_{ij}, are defined as:

$$q_{ij} = \frac{\left(1 + ||y_i - y_j||^2\right)^{-1}}{\sum_{k \neq l} \left(1 + ||y_k - y_l||^2\right)^{-1}}$$

The KL divergence (or cost function) between the joint probability distribution, P, in data space and of the low dimensional mapping, Q, to be minimised using gradient descent is:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

where p_{ii} and q_{ii} are set to zero.

The most important hyperparameters for tuning t-SNE are the perplexity, learning rate, and to a certain extent, early exaggeration. Perplexity can be thought of as setting the effective number of nearest neighbours each point is attracted to, which effectively sets σ for each data space point. The larger the value, the more non-local (global) structure will be retained in the projection. Lower values tend to generate smaller clumps of points. Perplexity is usually set to between 5 and 50, though is highly dependent on the dataset. The learning rate sets the step size for the gradient descent algorithm performed to minimise the KL divergence. Early exaggeration controls how tight clusters in the data space are in the embedding space.

3.3.9 Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection (UMAP; McInnes et al. 2018) is a non-linear dimensionality reduction technique that operates similarly to t-SNE. The main difference is how the data distributions in high and low dimensional space are defined and the nature of the cost function to minimise. The algorithm begins with approximating a manifold the data is assumed to (approximately) lie on. This is done by constructing a k-nearest neighbours graph for each data point, where each point is connected to its k-nearest neighbours. The kth nearest neighbour for each point determines the distance scale for that point's neighbourhood. Based on these local distance metrics, simplicial

sets are constructed. Simplicial sets are mathematical structures composed of simplices such that a 0-simplex corresponds to a point, a 1-simplex represents a line segment that connects two 0-simplices, a 2-simplex forms a triangle and a 3-simplex creates a tetrahedron, and so on. By combining these simplicial sets, a topological surface or manifold is defined. This surface captures the connectivity and relationships between data points in the high-dimensional space. The connections are weighted based on the distances between data points, resulting in a fuzzy topological representation that accounts for the flexibility in the data.

The lower dimensional space is defined similarly with the exception that distance is defined uniformly across the manifold using Euclidean geometry rather than varying based on the local neighbourhood of the data point. To find a low-dimensional representation that closely matches the topological structure of the original data, UMAP adjusts the layout of points in low-dimensional space by minimising a cost function given in the form of cross-entropy. The cross-entropy for UMAP measures the dissimilarity between the fuzzy topological structures of the high-dimensional data and the low-dimensional representation.

Parameters for the model include *n_components* (number of dimensions); and *n_neighbours* which controls the area of the local neighbourhood that UMAP looks at for each sample when building a manifold. With smaller values of *n_neighbours*, we focus on local structure, though with a risk of losing the bigger picture. Larger values, however, result in a broader view, at the risk of losing the finer structure within the data. *min_dist* controls the distance between data points. Lower values will result in clumpier embeddings, allowing you to see individual clusters more easily, while larger values enable you to see the broader topological structure. The *metric* parameter just represents the formula used to calculate the distance between points, the default is Euclidean.

3.3.10 Generative Topographic Mapping

Generative Topographic Mapping (GTM; Bishop et al. 1998) is a neural network-based manifold learning algorithm that computes a mapping between points in a low dimensional (often 2D) latent space into a higher dimensional data space such that the latent space representation reflects the data space distribution of data points. To do this, K

points in latent space are arranged in an equally spaced grid of nodes, $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K$, whose probability distribution is defined by delta functions.

$$p(\mathbf{x}) = \frac{1}{K} \sum_{i=1}^{K} \delta(\mathbf{x} - \mathbf{x}_i)$$

Each node $\mathbf{x}_i$ is mapped to a corresponding point $\mathbf{y}$ in data space using $\mathbf{y}(\mathbf{x}_i; \mathbf{W})$, where $\mathbf{W}$ is a matrix of parameters. This situation can be viewed as an L-dimensional non-Euclidean manifold confined within D-dimensional data space. The distribution of data points $\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_N$ is expected to be confined to the L dimensional manifold, though this is not true in reality. Therefore, one introduces a noise model such that the probability of observing a data point $\mathbf{t}$ in data space, given node $\mathbf{x}$ and parameterised by $\mathbf{W}$, follows a Gaussian distribution centred at $\mathbf{y}(\mathbf{x}; \mathbf{W})$ with a variance of $\frac{1}{\beta}$:

$$p(\mathbf{t}|\mathbf{x}, \mathbf{W}, \beta) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left[-\frac{\beta}{2}||\mathbf{y}(\mathbf{x}; \mathbf{W}) - \mathbf{t}||^2\right] \tag{3.4}$$

Taking into account all nodes, Equation 3.4 becomes:

$$p(\mathbf{t}|\mathbf{W}; \beta) = \int p(\mathbf{t}|\mathbf{x}; \mathbf{W}; \beta) p(\mathbf{x}) d\mathbf{x} = \frac{1}{K} \sum_{i=1}^{K} p(\mathbf{t}|\mathbf{x}_i; \mathbf{W}; \beta)$$

To find $\mathbf{W}$ and β the maximisation of the log-likelihood is required, given by:

$$L(\mathbf{W}; \beta) = \sum_{n=1}^{N} \ln\left(\frac{1}{K} \sum_{i=1}^{K} p(\mathbf{t}_n|\mathbf{x}_i; \mathbf{W}; \beta)\right)$$

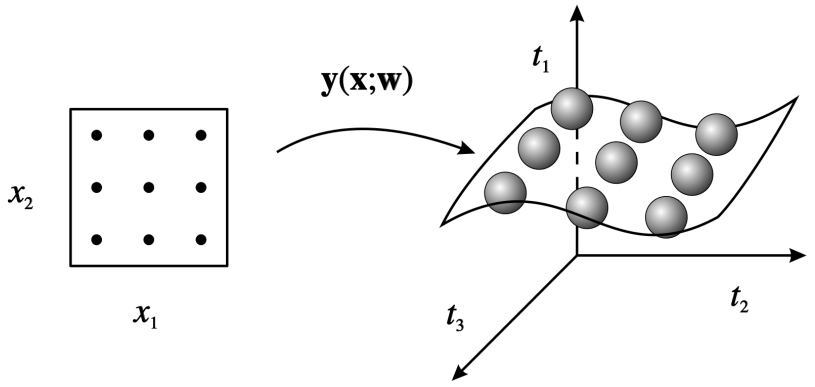


FIGURE 3.6: We consider a prior distribution $p(\mathbf{x})$ consisting of a superposition of delta functions, located at the nodes of a regular grid in latent space. Each node $\mathbf{x}_i$ is mapped to a corresponding point $\mathbf{y}(\mathbf{x}_i; \mathbf{W})$ in data space, and forms the centre of a corresponding Gaussian distribution Bishop et al. 1998.

The situation is depicted in Figure 3.6. The mapping function $\mathbf{y}(\mathbf{x}; \mathbf{W})$ takes the form of $y(\mathbf{x}; \mathbf{W}) = \mathbf{W}\phi(\mathbf{x})$, where elements of $\phi(\mathbf{x})$ consist of M radial basis functions $\phi_j(\mathbf{x})$, and $\mathbf{W}$ is a D×M matrix. The maximisation of the log-likelihood is performed using the Expectation Maximisation algorithm which iteratively updates the parameters $\mathbf{W}$ and β of the model to maximise the likelihood of the observed data. In each iteration, it calculates the responsibilities of each Gaussian component for every data point, defined as:

$$R_{in}(\mathbf{W}_{old}, \beta_{old}) = p(\mathbf{x}_i|\mathbf{t}_n, \mathbf{W}_{old}, \beta_{old})$$

It then updates $\mathbf{W}$ based on these responsibilities, and then updates β accordingly. The 2D representation of the higher dimensional data is the unfolded manifold upon which individual data points are projected in a location reflective of nodes (or Gaussians in data space) most responsible for them — node responsibility map.

Hyperparameters of GTM comprise the square root (sqrt) of the number of GTM nodes, k, sqrt of the number of RBF centres, m, the RBF width factor (RBF variance), s, and the regularisation coefficient. *regul*. These are associated with the Python implementation of GTM, *ugtm* (Gaspar, 2018), used within this work.

3.4 Data Preprocessing Techniques

Preparing data for input into ML algorithms, or pre-processing, is typically required before its usage. Handling missing data and feature scaling are two of the most crucial tasks for the production of optimal ML models.

3.4.1 Missing Data Handling

Most machine learning algorithms assume complete information for all features, no missing values (or NaNs) (Soley-Bori, 2013). Missing values in datasets derived from astronomical data such as time-series and astrometric data may arise due to insufficient data points in light curves for a feature to be extracted, data unavailability (e.g., parallax information due to short observational baseline), and erroneous data. There are several common approaches to addressing such issues. One may drop the entire column should it contain a missing value, however, valuable information may be discarded. Imputation is another strategy, in which a value is inserted based on the values in the remainder of the column (e.g., column mean) or other relevant columns. An extension to this approach involves adding another column indicating rows of imputed values. This works under the assumption that a missing value is informative for inference. Mean and K Nearest Neighbour imputation methods are methods adopted in this research.

Imputation

Imputing the mean of the column values for missing data in that column is simple and parameter-free. It is appropriate under the assumption that the data is normally distributed and most observations are around the mean anyway. One must be careful of its use, however, as this method ignores relationships between features and reduces the variance of the variable thereby introducing bias to the model.

The K Nearest Neighbour imputation method (Troyanskaya et al., 2001) operates within feature space, the N-dimensional space defined by the N dataset features/variables. For each dataset example, each missing feature is imputed using the values from the K nearest (based upon some distance metric, typically Euclidean) neighbours in feature space where that feature value is present. The imputed value will be either the uniform

or weighted-by-distance average feature value for those neighbours. The value of K is user-defined and set to 5 as default.

3.4.2 Feature scaling and transformation

Feature scaling is used to normalise the range of values of features, while transformation involves altering the distribution or scale of the data to suit machine learning algorithms. Feature scaling largely comprises normalisation and standardisation. Normalisation, also referred to as min-max scaling, consists of rescaling the range of values to lie in the range [0, 1] using the formula $x' = (x - min(x))/(max(x) - min(x))$. One may choose to normalise to a range in any arbitrary interval [a, b], in which case we use: $x' = a + ((x - min(x))(b - a))/(max(x) - min(x))$. Standardisation results in features with zero mean and unit variance. With the mean $\bar{x}$ and standard deviation σ of a feature, the following formula is used for standardisation $x' = (x - \bar{x})/\sigma$. Taking the logarithm (base 10) of the data, or log transform, is useful in reducing the skewness of heavily skewed distributions.

Normalisation is useful when the data distribution does not follow a Gaussian distribution. Neural Network algorithms can be sensitive to the scale of input features, preferring data on a 0 to 1 or -1 to 1 scale, otherwise, convergence of the algorithm to error minimum may be inhibited. Standardisation is often used where the data follows a Gaussian distribution. Standardisation does not have a bounding range, so, outliers are not affected by standardisation. Where such methods are not required are Decision Tree-based algorithms (Section 3.3.1). Such algorithms work feature by feature in their decision-making process rather than within multi-dimensional feature space.

3.5 Feature Selection Methods

As you add dimensions (features) you rapidly increase the minimum amount of samples required to adequately represent all combinations of feature values in your dataset. Increasing the dimensionality increases the complexity of the model whilst also causing the model to become increasingly dependent on the training set, thus leading to overfitting. Selecting the features most informative for our task enables ML algorithms to train faster, reduces complexity allowing for easier interpretation, reduces overfitting,

and can improve model accuracy for the right subset of features. Feature selection techniques may be grouped into filter methods that measure the relevance of features by their correlation with the dependent variable; and wrapper methods, that examine the usefulness of a subset of features by training a given model on them. The following describes several methods used in the work.

To identify the optimal feature subset, the Variance Inflation Factor (VIF; Vu et al. 2015), the one-way Analysis Of Variance (ANOVA; Quirk 2012), and the mutual information score (Quirk, 2012) methods are examined from the filter feature selection family. From the wrapper method family, the forward feature selection method was chosen.

3.5.1 Forward Feature Selection

Forward feature selection (FFS) is an iterative method starting with a model with no features. With each iteration, we add a feature, the one that produced the greatest increase in a performance metric as measured on a validation set. The process continues until no further performance increase is measured. The set of selected features may differ based on the choice of machine learning algorithm. Different algorithms often work best with distinct subsets of features, and the method can adapt to these individual requirements.

3.5.2 Variance Inflation Factor (VIF)

VIF (VIF; Vu et al. 2015) is a method used to detect multicollinearity - the existence of a linear relationship between two or more explanatory (independent) variables. It measures how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are uncorrelated. It is found by regressing each independent variable on the remaining independent variables to assess the degree to which it is explained by the remaining variables. VIF is given by:

$$VIF = \frac{1}{1 - R^2} \tag{3.5}$$

where

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{3.6}$$

where SS_{res} is the sum of squared residuals to the line of best fit in a linear regression model, while SS_{tot} is the sum of squared residuals to the average value. One uses this selection method by iteratively removing features with the highest VIF and recalculating the metric. A VIF equal to 1 represents the absence of multicollinearity, while the effects of multicollinearity increase with increasing VIF. While it is desirable to have VIF as close to 1 as possible, this generally leads to the removal of variables that have a high positive impact on model performance if we are not careful with our implementation of the technique. One must be careful to ensure the feature calculation is present in some form within the remaining features to maintain the associated information. VIF is particularly beneficial when dealing with feature redundancy that may arise when two or more features describe the same characteristic.

3.5.3 One-way ANOVA

One-way ANOVA (ANOVA; Quirk 2012) compares the mean value of a variable for each of three or more groups. It determines if any of those means are statistically significantly different from each other. The null hypothesis states that there is no statistically significant difference between any two group means:

$$H_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4 = ... = \mu_k \tag{3.7}$$

where μ is a group mean and k is the number of groups. The alternative hypothesis states that at least one of the groups is statistically significantly different from another at a significance threshold of 5%. This statistic was used to identify the significance of each feature ordered by p-value. A given algorithm was then trained using the top x% of the most significant features and the model cross-validation performance was recorded. This step was repeated, increasing the values of x in 5% increments from 5% to 95%, to arrive at a subset of features where model performance was strongest. This method is akin to forward feature selection, though with features added based on a statistical test rather than overall model performance. The motivation for the usage of one-way

ANOVA lies in its goal to select a set of features that hold significant importance in differentiating between classes.

3.5.4 Mutual information

Mutual information (MI; Quirk 2012) is the application of information gain (typically used in the construction of decision trees) to feature selection. The MI score measures the degree to which two variables are related. A score of zero is produced if the two variables are independent, and higher values for higher dependencies. For two jointly discrete random variables x and y, MI takes the form:

$$Mutual\ Information = \sum_{x \in X} \sum_{y \in Y} p(x, y) \ln \left[\frac{p(x, y)}{p(x)p(y)} \right] \tag{3.8}$$

We make use of the *scikit-learn* implementation, which uses a nearest neighbour method instead of binning to handle cases where the independent variable (feature), x, is continuous, assuming a discrete target, y, (see Ross 2014). Under the MI feature selection protocol, the most performant features were identified in the same way as for one-way ANOVA, resulting in slight variations in the optimal subset of features for each algorithm. In a similar fashion to one-way ANOVA, MI aims to select features most crucial for class distinction. However, MI quantifies the information shared between features and the outcome, thereby unveiling non-linear, intricate relationships.

3.6 Data Augmentation

A classification dataset with skewed class proportions is said to be imbalanced. Class imbalance can skew model predictions, with classifiers favouring the majority class while neglecting minority ones. Specifically, ML algorithms are usually designed to maximise accuracy (fraction of correctly predicted examples). So for a severe class imbalance of say 95:5, an algorithm may be inclined to classify everything as the majority class and achieve a 95% accuracy. Data augmentation techniques, such as random oversampling, undersampling, or synthetic data generation, aim to reduce these adverse effects by balancing the class distribution. The algorithms are then more likely to produce predictive models better capable of accurate predictions on real-world data. The methods

adopted in this research are majority class undersampling, class weighting and synthetic oversampling using the ADASYN algorithm.

3.6.1 Random Undersampling

Randomly selecting a subset of the majority class is a fast and easy way to balance a dataset. It is particularly effective in combination with using data augmentation techniques to increase the number of examples of minority classes, that way the amount of data lost from the majority class can be minimised. Random undersampling can be performed with or without replacement. We implement this without replacement.

3.6.2 ADASYN

Adaptive Synthetic (ADASYN; Haibo et al. 2008), a minority class oversampling technique, is a variation of the Synthetic Minority Over-sampling Technique (SMOTE; Chawla et al. 2002). SMOTE works by selecting a random example from the k nearest neighbours in feature space of a randomly chosen example from the minority class (or class of choice); draws a line in this feature space between the examples and generates a new sample at a random point along that line. The ADASYN adaptation generates more synthetic examples in regions of feature space where the density of minority examples is low, and fewer or none where the density is high. It does this by identifying regions in feature space where minority class instances are sparse, calculating the local density of minority instances around each sample as well as the imbalance ratio of minority to majority class examples in those regions. It then generates synthetic examples in low-density regions, prioritising regions where the imbalance is higher. The samples are generated, as with SMOTE, by interpolating between minority class examples. The result is that more synthetic data is generated for minority class samples that are harder to learn compared to those where many examples are available, thereby making it easier to learn the minority class properties.

3.6.3 Class Weighting

Rather than augmenting the dataset, one may modify the algorithm to account for skewed class distributions by giving different weights to each class depending on their

dataset prevalence. The difference in weights influences the classification during the training phase. The goal is to penalise the miss-classification of the minority class by setting a higher class weight, while at the same time reducing the majority class weight. Weightings are applied within the cost function for each algorithm such that the miss-classification of a minority class example leads to a greater cost penalty than for a majority class example.

Most of the scikit-learn classifiers have an in-built parameter *class weight* which helps us optimise the scoring for the minority class. By default, it is set to *None*, i.e., equal weights are assigned to each class. By setting this to *balanced*, the model automatically assigns class weights inversely proportional to their proportions in the dataset. The formula to calculate this is $w_j = n\_samples/(n\_classes \times n\_samples_j)$, where, w_j, $n\_samples$, $n\_classes$, and $n\_samples_j$ are the weight for each class, j, total number of dataset examples, the number of different classes, and the total number of examples of class j. Alternatively, manually setting the class weights is an option.

3.7 Model Evaluation and Hyperparameter Tuning

Performance metrics rely on the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). To compute these counts, one must establish the positive class, representing the class of interest (e.g., one of the CV classes), and the negative class, encompassing all other classes.

3.7.1 Confusion Matrix

Frequently, the counts of TP, TN, FP, and FN are organised in an N $\times$ N table referred to as a *confusion matrix*, with N signifying the number of classes. This matrix provides a straightforward means to view the quantities of TPs, TNs, FPs, and FNs. These values are used to calculate the class-specific precision, recall, and F1-score, as well as the balanced accuracy and the area under the curve of the receiver operating characteristic.

3.7.2 Precision, Recall, and F1-score

The precision is defined as the fraction of examples our model predicted as belonging to the positive class that does belong to this class: $TP/(TP+FP)$. In other words, it tells us how much we can trust our model's predictions of the positive class. The recall is the fraction of examples of the positive class that our model correctly predicted as belonging to this class: $TP/(TP+FN)$. This metric assesses the model's ability to identify all members of the positive class. The F1-score is the harmonic mean of precision and recall for our positive class and is useful in finding the best trade-off between these quantities. The highest possible value of the F1-score is 1 (100%), indicating perfect precision and recall, the lowest possible value (0) relates to a score of 0 for either precision or recall.

3.7.3 Accuracy and Balanced Accuracy

The accuracy is the fraction of all examples whose class was correctly predicted by the model. For binary classification, this is $(TP+TN)/(TP+TN+FP+FN)$, for a multi-class situation we sum the number of true positives for each class and divide by the total number of examples. The accuracy returns an overall measure of the model's predictive capability. Should we only be concerned with assigning the most number of examples to their correct class, accuracy is a good metric. However, under this metric, high classification errors for classes with few examples to their name will be hidden. Therefore, should we be concerned with finding a model which has a strong classification performance across all classes, we may use 'balanced accuracy' which can account for this class imbalance. This is calculated as the arithmetic mean of the recalls for each class.

3.7.4 Area under the Curve of the Receiver Operating Characteristic (AUC)

The Receiver Operating Characteristic (ROC) curve offers a visual representation of the trade-off between sample purity and completeness. It plots the true positive rate (TPR), also known as recall, against the false positive rate (FPR). The FPR represents the fraction of examples incorrectly classified as belonging to the positive class, calculated as $FP/(TN+FP)$. This curve is generated by varying the threshold probability used to

determine positive classifications for each example. In detail, ML algorithms provide a class probability score for each example, and a threshold is applied to classify examples as positive or negative. The ROC curve showcases the performance of the TPR and FPR as this probability threshold is continuously adjusted. This tool is valuable for selecting an appropriate threshold that aligns with the desired balance between purity and completeness, depending on the specific research objectives. In classification tasks, the goal is to maximise TPR while minimising FPR. An area under the curve (AUC) value of 1 indicates a perfect model that correctly assigns class predictions for all examples. An AUC of 0.5 signifies a model no better than random guessing, while an AUC of 0 implies incorrect predictions for all examples. Although ROC curves are typically associated with binary classification, in the case of multi-class models, they are generated using a one-versus-rest approach. This entails designating one class as the positive class and the remaining classes as the negative class to produce separate curves for each class.

3.7.5 McNemar's Test

While performance metrics can be used to assess test set performance differences between two classifiers, the McNemar's test can be utilised to judge significant differences between their predictions on the test set and in some sense whether any performance difference is significant. The null hypothesis states that the classifiers disagree in their class predictions to the same amount. Should this be rejected, the alternative hypothesis implies there is evidence they disagree in different ways. The test statistic is calculated in the following way:

$$statistic = \frac{(Yes/No - No/Yes)^2}{Yes/No + No/Yes} \tag{3.9}$$

where Yes/No is the number of test instances that classifier 1 got correct and classifier 2 got incorrect, while No/Yes describes the opposite of this. The test statistic follows a chi-squared distribution with one degree of freedom. The test is usually administered in a binary classification setting, however, under the multi-class case, the correct and incorrect classifications are performed for each class.

3.7.6 Hyperparameter Tuning

Several of the aforementioned metrics may be used to tune hyperparameters associated with each of the machine-learning algorithms. Adjusting the hyperparameters under which ML models are trained can help to mitigate underfitting and overfitting. Underfitting refers to the situation where the model has not captured the relationships or intricacies within the dataset during training and performs poorly on unseen data. Overfitting refers to the situation where the model has not only learnt patterns in the dataset but also the noise meaning that it achieves a high accuracy on the training set but does poorly on the validation and test sets (or in production).

Cross Validation

The validation set is used to tune algorithm hyperparameters that control how a model is trained, while the test set is held back, taking no part in the training and model-tuning process. Should the size of the dataset be insufficient for a separate validation set, for example, in cases where minority class examples are few, stratified k-fold cross-validation may be used. This involves splitting the training set into k separate subsets (or folds) in a stratified manner — each fold contains the same class proportions as the overall training set. A model is trained on k-1 folds and evaluated, based on a given metric, on the remaining fold (validation fold); this step is repeated until each fold has partaken in the validation process. The metric scores for each of the k models are mean averaged to produce a cross-validation score. This technique allows an adequately sized training set to be maintained.

The scikit-learn Python package offers automated methods of searching the hyperparameter space for the best cross-validation score. The associated set of hyperparameters will be those which produce the best model. Given a set of hyperparameters with associated test values, *GridSearchCV* considers all parameter combinations to identify the optimal combination. This type of search is exhaustive though can be time-consuming especially when the grid of parameters is large. *RandomizedSearchCV* on the other hand, is less exhaustive but also less time intensive. It randomly samples a user-specified number of parameter combinations from the distribution of all those possible in the parameter grid.

Chapter 4

Source Classification

Having provided a comprehensive overview of CVs and explored the machine learning techniques used in this study, I will discuss how surveys have played a crucial role in expanding our understanding of CVs and how machine learning source classification can aid in this process through the efficient handling of vast amounts of survey data. The chapter concludes by laying out the research problem in light of the above and provides a brief overview of the chapters that follow.

4.1 Impact of Time Domain Surveys

Wide field time domain surveys have been responsible for the discovery of many of the currently known list of CVs. For example, CRTS obtains unfiltered images of 30,000 square degrees of the sky with three ground-based telescopes at a $\sim$ 2 week cadence. Drake et al. (2014) analysed over 5 years of CRTS data to report the discovery of 705 new CV candidates. The Gaia space mission (Gaia-Collaboration et al., 2016) observes the whole sky (including the Galactic plane) from the L2 Lagrange point at a cadence of 2 to 4 weeks. It publishes detections of new transients via Gaia Science Alerts (GSA; Hodgkin et al. 2021). Over 2,500 confirmed or candidate CVs reside amongst GSA. Rare examples can also be picked out; by eyeballing GSA lightcurves, Gaia14aae became the first fully eclipsing AM CVn to be discovered (Campbell et al., 2015). At higher cadences, ASAS-SN uses multiple telescopes to survey the entire visible sky every night down to about 18th magnitude. As a consequence, the ASAS-SN alerts page (Shappee et al.,

2014) contains over a thousand CVs or candidates within its list, with new candidates reported via an Astronomer's Telegram (e.g, Jayasinghe et al. 2020; Prieto et al. 2013). ZTF observes fields north of $\delta = -31$ deg every 2 to 3 days in two bands, ZTF-g and ZTF-r, to depths of 20.8 and 20.6 magnitudes, respectively. Szkody et al. (2020, 2021) filtered the ZTF transient alerts by looking for point sources with g-r colour < 0.6 and a magnitude change $\Delta m >= 2$ within a timescale of 2 days in the g band. This resulted in a total of 701 known or candidate CVs over two years of its implementation that typically displayed dwarf nova outbursts and changes in accretion state. An extension of this filter-based approach was performed by van Roestel et al. (2021) to uncover nine new outbursting AM CVns from ZTF. These discoveries have helped fill gaps in our current knowledge, for example, constantly evolving models are being developed attempting to explain the diversity of dwarf nova outbursts based on the disk instability model (Kotko et al., 2012; Hameury, 2020); Knigge et al. (2011) was able to construct semi-empirical models for the evolution of CVs based on donor star masses and radii. However, discoveries also uncover new gaps, such as the detection of pulsed X-rays in two AM CVns that not only raises the question of magnetically controlled accretion in AM CVns but has implications for their evolutionary timescales (Maccarone et al., 2023).

4.2 Machine Learning in Astrophysics

ML is becoming ever more present in the field of astrophysical source identification and classification due to its necessity in characterising the vast numbers of transient events detected by time domain surveys. The following is a selection of pertinent examples of its usage in the literature. The ZTF Source Classification Project (van Roestel et al., 2021) is a framework that aims to group transients based on both variability types and transient classes. Variability types (or phenomenological classes) are comprised of irregular, periodic, flaring and eclipsing variability (amongst others). Transient classes include AGN, YSOs, several variable star classes, and binary stars within which CVs are encompassed amongst other classes of binary. For each class, a classifier is trained to distinguish between that transient and the remainder (one versus rest). Both a 'dmdt' mapping of g band light curves and statistical, periodicity and percentile-based features were used within a CNN and XGBoost. Each of the models (one for each class)

performed well when analysed against a test set, with an accuracy of no lower than 0.85 and 0.96 for CNN and XGBoost models, respectively. Performance on the test set says little about the performance on the full corpus of ZTF light curves, which Mahabal et al. (2017) examined. For RR Lyrae (a pulsating variable), 89% (2102) of the 34 million light curves that were classified as RR Lyrae (with a probability $> 90\%$), did belong to the class after visual inspection. However, the classification of YSOs was less successful, with only 26% of those identified belonging to the class.

Rimoldini et al. (2022) aimed to classify all sources detected by Gaia as variable into one of 24 variability types/classes using Random Forest and XGBoost. The work provides candidate source lists to a wide variety of research groups focusing and different classes of transient. The classes included several types of pulsating stars, eclipsing binaries, ellipsoidal variables, spotted stars, eruptive and cataclysmic phenomena, stochastic variations of AGNs, microlensing events, and planetary transits. Input for these algorithms were basic statistics, photometric colours, astrometric parameters, periodicity indicators, and combinations thereof, all extracted from the photometric time-series in the Gaia G, BP, and RP bands. The training and test sets were formed by cross-matching sources of known variability type from literature with Gaia DR3 variable sources. Several different types of classifiers (models) were trained - multi-class, binary (one versus rest), and hierarchical classifiers. They were implemented to provide classifications for 12.4 million Gaia DR3 variable sources. Classifications were assigned to each source based on combining the posterior probabilities (probability of class belonging) output by each classifier. For the CV class in particular, 7,306 sources were identified as candidates, 233 of which are known to be CVs, though over 1200 known CVs were misclassified as belonging to other variability types.

Neira et al. (2020) tested an MLP, RF, and SVM to classify 4869 CRTS light curves into 8 transient classes (AGN, Blazar, CV, flare stars, high proper motion stars, supernovae, non-transients, and other). Inputs for the algorithms were traditional statistical, magnitude-based (amplitude, maximum slope, ...) and percentile-based, along with coefficients used for fitting polynomials to the light curves. RF performed the best though with precision and recall scores of 49% and 70%, respectively, when averaged over all classes. For the CV class, a precision and recall of 74 and 76%, respectively, were achieved.

Sun et al. (2021) searched for CVs within Data Release 6 (DR6) of the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST; Cui et al. 2012) survey containing nearly 10 million low-resolution spectra. The methodology was both machine learning and visual inspection-based. The machine learning phase involved the use of KNN to separate sources with Hα emission from those without, as broad Hα emission is a particular feature of hydrogen CV spectra. A reduced dimensionality representation of the flux measurements around the Hα wavelengths using UMAP was used as input for KNN. This reduced the 10 million strong list to 169,509 Hα emitting sources that each underwent a process of visual inspection. This involved comparison with stellar spectral templates for further filtering and inspection for spectral features characteristic of CVs. This process resulted in 323 CVs or candidate CVs being identified, of which 52 are new candidates.

Following on from this work, Hu et al. (2021) explored DR7 of LAMOST containing 10.6 million low-resolution spectra for CVs with ensemble learning algorithms. The dataset comprised 567 confirmed CV spectra from SDSS and LAMOST serving as positive instances, along with 20,000 LAMOST spectra categorised as non-CVs for the negative class. The input data consisted of standardised spectral flux measurements across 3,473 wavelength bins spanning from 4000 to 8900 Åfor each spectrum, where the flux value at a specific wavelength represented a distinct feature. Gradient Boosting algorithms XGBoost and Light Gradient Boosting Machine (LightGBM; Ke et al. 2017) produced the best-performing models with accuracy, precision, recall, and F1-scores all at least above 92% (reaching $\sim$ 99.7% accuracy). Feature importance scores revealed the importance of Balmer lines as well as HeII(4685 Å), and HeI(5876 Å). Implementing the LightGBM model on LAMOST-DR7 uncovered 255 CV candidates, 4 of which are new discoveries.

4.3 Research Problem

Non-ML filter-based approaches have demonstrated their effectiveness at CV identification (e.g., Szkody et al. 2020, 2021; van Roestel et al. 2021) though the requirement for human vetting is significant. ML-based photometric and spectroscopic approaches have also been somewhat effective, demonstrating that searches for CVs with ML is an active field of research. However, the ML-based approaches treat CVs as a broad class

of transient. While broad classification/identification of CVs is important, the ability to automatically group such targets into their respective subtypes, and/or identify rare subtypes is the ultimate goal. In the case of CVs, rarities include the ultrashort period (5–65 min) AM CVns (Solheim, 2010), magnetic CVs (Cropper, 1990; Patterson, 1994), and eclipsing sources from which accurate parameters can be derived (van Roestel et al., 2022; Wakamatsu et al., 2021; Hope & Copperwheat, 2019). Such a classifier/pipeline should dramatically reduce the requirement from human vetting, something that is expected to become all the more important as transient/variable source detection capabilities improve with time. The end goal of my research is to develop a machine learning pipeline capable of utilising survey data to identify/classify CVs on a more granular level, i.e., pick out the different subclasses. This is currently unexplored territory and especially important should we aim to serve research groups that focus on specific CV subtypes whose goal is to formulate an accurate picture of CV evolution and better understand the processes responsible for their variability.

To address the research gap, in Chapter 5, which consists of my first publication (Mistry et al., 2022), I embark on an exploration of ML techniques to unearth CVs within the confines of a low cadence survey that utilises data from the Gaia spacecraft (Gaia-Collaboration et al., 2016). The time-series photometry of transient/variable sources provided within the Gaia Science Alerts (Hodgkin et al., 2021) resource provides the input. The experience obtained within that work provides a platform for a more granular approach, the identification/classification of CVs from within the ZTF alert stream; the higher cadence provides the opportunity for subtype classification. This research forms Chapter 6, the contents of my second publication (Mistry et al., 2023). The results of Chapter 6 lead nicely onto an unsupervised learning approach for ZTF CVs in Chapter 7. Chapter 7 explores the high dimensional structure of ZTF CVs (using their light curve properties) with dimensionality reduction techniques (PCA, t-SNE, UMAP, and GTM). The thesis ends (Chapter 8) with a discussion of the key findings of my research efforts and avenues of future research concerning automated CV searches.

Chapter 5

Gaia exploration

5.1 Introduction

In this work, I describe the exploration of data generated by the Gaia spacecraft (Gaia-Collaboration et al., 2016) to identify new members of the CV population. Gaia is now recognised as a powerful tool for transient detection, with Gaia Science Alerts (GSA; Hodgkin et al. 2021) providing alerts of newly discovered transient sources at a current rate of ∼12 per day by repeatedly scanning the whole sky. The cadence of the associated light curves is dictated by the 'Gaia scanning law' (Gaia-Collaboration et al., 2016) — typically, a pair of observations separated by 106.5 minutes are separated by another pair two to four weeks later. The photometry is precise to 1% at G=13, and 3% at G=19. This resource therefore provides a stable platform from which to evaluate ML-based classification. In Section 5.2, I describe the classified transients of GSA; the methods used to extract relevant descriptive characteristics from their light curves; and the additional metadata gathered from the survey for each source. In Section 5.3, I describe how the resultant dataset was used to train several ML algorithms to perform a set of classification tasks, along with a description of how the resultant models can be evaluated. In Section 5.4, I detail the performance of each algorithm. Finally, I discuss the outcomes of the exploration of GSA along with a description of a pilot study involving spectroscopic classification to validate predictions made by the best-performing model (Section 5.5).

5.2 Dataset

5.2.1 Gaia alerts and EDR3

As of June 2021, close to 18,000 transient sources had been listed within the Gaia transient alerts stream [1]; just over 4,700 of which had been assigned class labels. The classifications are based upon human inspection of Gaia data in combination with the results of positional cross-matching with the Simbad (Wenger et al., 2000), NED and VSX databases[2], and YSO catalogues (see section 2.7.7 of Hodgkin et al. 2021) to identify already-confirmed transient or variable objects. This information is aided by the hourly parsing of 27 major transient survey websites for reported discoveries that also contain classification information, these include Transient Name Server (TNS)[3], CRTS, ASAS-SN and Astronomer's Telegrams[4]. Further details regarding the alerts filtering and classification process are contained in Hodgkin et al. (2021).

The process of training and validating machine learning models requires accurate class labels. Whilst the aforementioned process of class assignment can reliably provide this accuracy, an inspection of class labels for a sample of these sources was performed for a level of verification. Of the 2,713 supernovae, 2,530 are spectroscopically confirmed according to TNS, Astronomer's Telegrams contain details of spectroscopic classification for the remainder. Of the 613 Gaia-labelled CVs, 471 are associated with known/confirmed CVs according to the comments associated with the Gaia classifications. Comparison with VSX confirms this with either a confirmation of CV status or candidate status for the remainder through references to relevant research papers and Astronomer's Telegrams. Gaia's comments associated with sources labelled as AGN and YSO show 929 of the 940 transients labelled as AGN, and 184 of the 190 transients labelled as YSOs are associated with known/confirmed AGN and YSOs respectively. This was verified for a sample of these sources by examining records within TNS and associated links (e.g., Simbad). The remaining candidate AGN and YSOs were not further considered for this work.

[1] `http://gsaweb.ast.cam.ac.uk/alerts/alertsindex`

[2] The NASA/IPAC Extragalactic Database (NED) is funded by the National Aeronautics and Space Administration and operated by the California Institute of Technology. VSX is the International Variable Star Index database, operated at AAVSO, Cambridge, Massachusetts, USA.

[3] https://www.wis-tns.org/

[4] https://www.astronomerstelegram.org

The dataset is composed of features extracted from light curves of these classified targets within Gaia's alert stream along with their associated class labels. Supplementary data for these targets may be available within the database of Gaia Early Data Release 3 (EDR3; Lindegren et al. 2021; Riello et al. 2021) in the form of astrometric and further photometric data such as parallax, proper motion, and photometric colour provided by the low-resolution photometry (R=100) of blue and red photometers onboard Gaia. A coordinate cross-match with EDR3 provides this metadata for ∼45% of sources within the dataset. This metadata has also been incorporated as a set of supplementary features.

Of the 4,697 classified targets incorporated into the dataset, SNe account for 58% of classified targets, AGN make up 21%, CVs and YSOs constitute 13% and 3%, respectively, while microlensing, tidal disruption events, and various other classes account for the remainder.

The majority of GSA classifications come from dedicated spectroscopic follow-up from, for example, the Public ESO Spectroscopic Survey of Transient Objects (PESSTO; Smartt et al. 2015) that uses the new technology telescope (NTT; Wilson 1991) with optical and near-infrared spectrographs; and the Spectral Energy Distribution Machine (SEDM; Blagorodnova et al. 2018), an integral field unit spectrograph mounted on the Palomar 60-inch telescope, utilised by the Bright Transient Survey (BTS; Perley et al. 2020) for classification of extragalactic objects brighter than 19th magnitude. These are heavily biased towards supernova classification. The class fractions of classified targets are generally dictated by what has been chosen to be classified, with unusual or ambiguous examples often overlooked, and therefore it must be noted that these fractions may not be representative of the entire sample of GSA targets.

5.2.2 Light curve feature extraction

Quantitative characteristics (or features) were extracted from source light curves to describe their variability. These included simple statistical and periodicity-based features in Table 5.1 along with features obtainable from the feATURE eXTRACTOR FOR tIME sERIES (FEETS) package (Cabral et al., 2018), a selection of which are shown in Table 5.2. The FEETS package specialises in analysing astronomical time-series data, which often exhibit irregular sampling, seasonal gaps, differences in cadence, and

variability in the number of data points. It is designed to be robust against these challenges. For example, the Lomb-Scargle periodogram, a key tool for analysing unevenly sampled data, forms the basis for several FEETS features, such as those which measure the amplitude and phase of periodogram frequency components (and their harmonics). Their derivation involves calculating the light curve periodogram, finding the strongest periodic signal (frequency) and its harmonics, subtracting the model for that signal (including its harmonics) from the data, repeating for several iterations to identify several frequencies, and finally extracting the amplitude and phase for each frequency and its harmonics. This approach accounts for measurement uncertainties through weighting and discourages high-frequency artefacts (caused by noise or irregular sampling) by penalising frequencies above the Nyquist limit. The diversity of FEETS features captures various types of variability, not limited to periodic signals. For instance, *Eta_e* quantifies overall variability, where high *Eta_e* values indicate rapid, erratic changes, while low values reflect smoother variations. This feature helps distinguish between periodic, semi-periodic, and stochastic behaviours. By combining these diverse features, the FEETS package provides a comprehensive characterisation of light curves, facilitating more accurate classification of variability types and improving the performance of machine learning algorithms.

5.2.3 Supplementary features

Supplementary data (or metadata) from Gaia EDR3 relating to position, photometry and astrometry are incorporated as dataset features. Positional features consist of: right ascension, declination, Galactic (and ecliptic) longitude and latitude, along with associated errors. Photometric features encompass the mean flux from the red and blue photometers (BP and RP) as well as that from G band photometry; the associated mean magnitudes; colours (BP-RP, BP-G, G-RP) and associated errors. Proper motion and parallax (along with their errors) are included as astrometric features. A full list is displayed in Table 5.3, while further details are available within the Gaia EDR3 documentation[5]

[5] `https://gea.esac.esa.int/archive/documentation/GEDR3/Gaia_archive/chap_datamodel/sec_dm_main_tables/ssec_dm_gaia_source.html`

TABLE 5.1: Features extracted from light curves (without `feets` package)

Feature	Description
mean_mag	Mean of magnitudes
median_mag	Median of magnitudes
std_mag	Standard deviation of magnitudes
mad_mag	Median absolute deviation of magnitudes
min_mag	Minimum magnitude (maximum brightness)
max_mag	Maximum magnitude (minimum brightness)
n_obs	Number of observations
diff_min_mean	Difference between *min_mag* and *mean_mag*
diff_min_median	Difference between *min_mag* and *median_mag*
detected_time_diff	Time span of observations
n_peaks_rm_x_y	Number of observations within a rolling window of y observations that are brighter than x magnitudes of the median magnitude of that window (x = *1, 2, 3, 4, or 5*, y = *7*).
kurtosis	Kurtosis of the magnitudes
skew	Skewness of the magnitudes
pwr_max	Largest power value in the Lomb Scargle Periodogram
freq_pwr_max	Frequency corresponding to *pwr_max*
FalseAlarm_prob	Estimate of the false alarm probability given the height of the largest peak in the periodogram (see `https://docs.astropy.org/en/stable/api/astropy.timeseries.LombScargle.html#astropy.timeseries.LombScargle.false_alarm_probability`)

5.3 Method

5.3.1 Machine Learning algorithms

The dataset described above can be used to evaluate the ability of ML algorithms to identify CVs within GSA. The algorithms whose performances are evaluated are SciKit-Learn's (Pedregosa et al., 2011) Python implementation of Random Forest (RF; Breiman 2001), AdaBoost (ADB; Freund & Schapire 1997), K-Nearest neighbours (KNN; Zhang 2016), and Support Vector Machines (SVM; Cortes & Vapnik 1995). Also used are the Extreme Gradient Boosting (XGBoost) algorithm (Chen & Guestrin, 2016) and Keras (Chollet, 2021) implementation of an Artificial Neural Network (ANN) in the form of a Multi-Layer Perceptron — a fully connected multi-layer ANN (Kruse et al., 2022).

TABLE 5.2: A small selection of features available from the `feets` package. The full list is available at (`https://feets.readthedocs.io/en/latest/tutorial.html`) along with detailed explanations. Of the full list, only those requiring a magnitude and time, or just magnitude data, were implemented here.

Feature	Description
Amplitude	Half of the difference between the median of the maximum 5% and the median of the minimum 5% magnitudes
AndersonDarling	The Anderson-Darling test is a statistical test of whether a given sample of data is drawn from a given probability distribution (normal distribution)
Autocor_length	Cross-correlation of a signal with itself
Eta_e (η^e)	Variability index η is the ratio of the mean of the square of successive differences to the variance of data points.
*FluxPercentileRatioMid**X***	Ratio of centred flux percentile ranges. If $F_{5,95}$ is the difference between the 95th and 5th percentile of ordered magnitudes, then *FluxPercentileRatioMid**X*** $= F_{40,60}/F_{5,95}$, $F_{32.5,67.5}/F_{5,95}$, $F_{25,75}/F_{5,95}$, $F_{17.5,82.5}/F_{5,95}$, and $F_{10,90}/F_{5,95}$, for $\mathbf{X} = 20$, 35, 50, 65, and 80 respectively.
*Freq**i**_harmonics_amplitude_j*	cAmplitude of the jth harmonic of the **i**th frequency component of the Lomb Scargle Periodogram
Gskew	Median-of-magnitudes based measure of the skew
LinearTrend	Slope of a linear fit to the light-curve
MaxSlope	Maximum absolute magnitude slope between two consecutive observations
Meanvariance	Ratio of the standard deviation to the mean magnitude
PairSlopeTrend	Considering the last 30 (time-sorted) measurements of source magnitude, the fraction of increasing first differences minus the fraction of decreasing first differences
PeriodLS	Period corresponding to frequency of maximum power in the Lomb Scargle Periodogram
PercentAmplitude	Largest percentage difference between either the max or min magnitude and the median
Psi_eta	η^e index calculated from the phase-folded light curve
SmallKurtosis	Small sample kurtosis of the magnitudes

TABLE 5.3: Supplementary data from Gaia EDR3 incorporated as dataset features (see subsection 5.2.3)

Feature	Description
ra, dec, ra_error, dec_error	Right ascension, declination, and associated standard errors
l, b	Galactic longitude and Galactic latitiude
ecl_lon, ecl_lat	Ecliptic longitude and Ecliptic latitude
bp_rp, bp_g, g_rp	BP-RP, BP-G, and G-RP colours
*phot_**X**_mean_flux*	Mean flux in the G, integrated BP, or integrated RP bands — corresponding to ***X*** = *g, bp, or rp* respectively
*phot_**X**_mean_flux_error*	Error on the mean flux in the **X** band. Standard deviation of the **X**-band fluxes divided by sqrt of the number of observations (data points)
*phot_**X**_mean_flux_over_error*	Mean flux in the **X** band divided by its error
*phot_**X**_mean_mag*	Mean magnitude in the G, integrated BP, or integrated RP bands — corresponding to ***X*** = *g, bp, or rp* respectively
pseudocolour, pseudocolour_error	The astrometrically estimated effective wavenumber of the photon flux distribution in the astrometric G band, measured in $\mu^{-1}m$, and standard error of pseudocolour
parallax, parallax_error	Gaia parallax in milliarcseconds (mas) and standard error
parallax_over_error	Parallax divided by its standard error
pm, pmra, pmdec	Total proper motion, and proper motion in the right ascension and declination directions (mas/year)
pmra_error, pmdec_error	Standard error of the proper motion in right ascension and declination directions (mas/year)
ruwe	renormalised unit weight error: expected to be around 1.0 for sources where the single-star model provides a good fit to the astrometric observations. A value significantly greater than 1.0 (say, > 1.4) could indicate that the source is non-single or otherwise problematic for the astrometric solution

5.3.2 Fine Tuning

To control how the algorithms learn from the dataset to generate predictive models, their hyperparameters must be adjusted/tuned in such a manner as to improve model performance. The hyperparameters explored for each algorithm are given in Table 5.4.

TABLE 5.4: The hyperparameters explored for each ML algorithm.

RF Hyperparameters	**Description**
n_estimators	Number of Decision Trees
max_features	maximum number of features provided to each tree
max_depth	maximum number of binary split levels in each tree
ADB Hyperparameters	
n_estimators	Same as for RF
learning_rate	Weight assigned to each classifier at each boosting iteration. This determines the impact of each tree on the final outcome.
max_depth	Same as for RF
XGBoost Hyperparameters	
n_estimators	Same as for RF
min_child_weight	Minimum sum of weights of all observations in a child node
gamma	Nodes are split only when there is a reduction in the error defined by a loss function. Gamma specifies the minimum loss reduction required to make a split
subsample	Fraction of examples to be randomly sampled for each tree
colsample_bytree	Similar to *max_features* in Random Forest
max_depth	Same as for RF
SVM Hyperparameters	
Kernel	see text: 'Radial Basis Function (RBF)'
Kernel Coefficient (γ)	Defines how far the influence of a single training example reaches, where the values can be seen as the inverse of the radius of influence.
Error Penalty (C)	Controls the cost of miss-classification on the training data. Small C = soft margin, large C = hard margin.
KNN Hyperparameters	
n_neighbors	Number of nearest neighbours to use
MLP Hyperparameters	
learning_rate	Controls how much to change the model in response to the error each time the model weights are updated.
Number of Hidden Layers	Number of hidden layers
Number of neurons	Number of units (neurons) within a given hidden layer.
Activation function	Converts the output of a neuron into a form that serves as input for the next. Used to introduce non-linearity to a network.

5.3.3 Classification tasks

The classes assigned by the Gaia team are not mutually exclusive. For example, quasi-stellar objects (QSOs) are extremely luminous AGN. From the Gaia-assigned classes, many variations of class grouping could be put forward for ML classification algorithms to distinguish. Two such groupings are defined by the following classification tasks, listed as transient class followed by the number of dataset samples in brackets:

1. **Binary classification** — CV (613) or not CV (4,084).
2. **4 class classification** — this comprises the most populous transient types in the dataset: AGN (which includes QSOs and BL Lac) as a single class (929), CVs (613), all different supernova types (SNe; 2,713) and Young Stellar Objects (YSOs; 184).

The tasks are assigned to the ML algorithms and their performance is evaluated. The classification tasks were first performed with both the light curve extracted features and supplementary features. However, between 58% and 90% of data is missing for supplementary features. This was either due to unsuccessful cross-matching of targets with EDR3 — cross-matching was unsuccessful for 90% of supernovae, 23% of CVs, and <1% of AGN and YSOs respectively — or certain metadata not being available where cross-matching was successful. For example, parallax measurements may not be available if the target is too faint or distant for an accurate measurement. Therefore, I felt it necessary to also perform classification tasks with light curve extracted features alone. These implementations can then be compared with other works where classification has been performed using light curve-derived features alone.

5.3.4 Data pre-processing

Prior to ingestion into ML algorithms, the associated datasets require some level of preparation. Examples within each task-specific dataset contain missing data for several features. The strategy employed here is to replace missing values with the mean value of the feature column (mean imputation; Khan et al. 2018). Feature scaling is employed for all except the ensemble learning algorithms (i.e., RF, ADB, and XGBoost) so that features with a larger range of values do not impart more influence on the model during

training and for faster convergence to error minimum for Gradient Descent algorithms. I standardised the data to achieve zero mean and unit variance (or equivalently, standard deviation; Muhammad Ali & Faraj 2014).

5.3.5 Train-test split

Training and evaluation of an ML model requires a separate training and test set. The algorithms are trained on the training set to generate a model to be evaluated on the test set. The task-specific datasets are split 50/50 into a training set and test set in a stratified manner — the same proportion of each class is represented in each of the training and testing sets. The split is performed before the pre-processing (imputation and feature scaling) stages to avoid information from the test set being present within the training set (data leakage) and yielding extremely biased results on model performance. A validation set was obtained via cross-validation, as described in the following subsection.

5.3.6 Optimal Hyperparameter Search

Manually testing all hyperparameter combinations to find the optimal set is computationally infeasible, particularly when using cross-validation. To address this, for the RF, ADB, XGBoost, KNN, and SVM algorithms, the *GridSearchCV* and *RandomizedSearchCV* functions from Scikit-learn's *model_selection* package (Pedregosa et al., 2011) were employed. These functions systematically explore predefined hyperparameter combinations and perform cross-validation for each. The optimal set of hyperparameters for a given algorithm is determined as the one achieving the highest balanced accuracy cross-validation score — representing the average balanced accuracy across validation splits. A 10-fold cross-validation approach was used (i.e., nine splits for training and one for validation obtained from the training set).

For the ANN, a manual tuning approach was adopted. The training set was divided into 90% for training and 10% for validation in a stratified manner (mirroring the validation split proportions for the non-ANN algorithms). All chosen hyperparameter sets were trained for a fixed, large number of epochs (1,000), but the training was halted early based on the validation set balanced accuracy to avoid overfitting. The model from the epoch that achieved the highest validation balanced accuracy was saved for further

evaluation. This approach ensured that the model did not overfit to training set noise and maintained good generalisation to unseen data. The evaluation of validation performance for each hyperparameter combination identified the optimal hyperparameters for the ANN algorithm.

5.4 Results

Tables 5.5 and 5.6 show the model evaluation scores for the binary and 4 class classification tasks. The scores for models trained with both light curve extracted and supplementary features (full feature models) are shown without brackets, while the scores for models trained with light curve extracted features alone (light curve only models) are within brackets. The scores shown are the accuracy, balanced accuracy, and with respect to the CV class, the precision, recall, and F1-score. The choice of best-performing model is based on the F1-score for the CV class. This metric was chosen as it considers both the need to minimise false positives (FPs), which is important for the efficient use of telescope time for target follow-up, and a requirement to minimise false negatives (FNs).

5.4.1 Binary classification

5.4.1.1 Full feature model

The best performing binary task full feature model was XGBoost, trained with 150 Decision Trees at a learning rate of 0.1 and maximum tree depth of 6. The model outperformed each of the others with an F1-score of 84%, the AdaBoost and Random Forest implementations follow closely behind (81-83%). There is though little difference between the top two models; use of the McNemar's test to compare the XGBoost and AdaBoost models shows both classifiers make errors in much the same proportions (for $\alpha = 0.05$; $p = 0.175$). The confusion matrix (top panel of Figure 5.1) indicates 69 of the 307 CVs in the test set were miss-classified by the XGBoost model, while of the 258 examples predicted as CVs only 20 were not. The corresponding ROC curve is plotted in the top panel of Figure 5.2, with an AUC score of 0.975. The importance of each feature for a given model can be given by the feature importance scores. The 20 features with the largest effect on the model's predictive accuracy are plotted in the top panel

TABLE 5.5: Binary task classification scores for ML models as measured on the test set. Scores without brackets relate to models using both light curve and supplementary features, while those in brackets are for models that used only light curve extracted features. Random Forest was implemented with 100, 250, 750, and 1000 trees denoted by RF then the number of trees; other abbreviations are ADA – AdaBoost, MLP – Multi-Layer Perceptron, KNN – K Nearest Neighbours and SVM – Support Vector Machine

Model	Accuracy	Balanced Accuracy	CV Precision	CV Recall	CV F1-score
RF100	0.955	0.870	0.88	0.76	0.81
	(0.938)	(0.824)	(0.82)	(0.67)	(0.74)
RF250	0.955	0.870	0.89	0.75	0.81
	(0.939)	(0.829)	(0.82)	(0.68)	(0.74)
RF500	0.955	0.868	0.89	0.85	0.81
	(0.937)	(0.827)	(0.81)	(0.68)	(0.74)
RF750	0.955	0.867	0.89	0.75	0.81
	(0.937)	(0.826)	(0.81)	(0.67)	(0.74)
RF1000	0.956	0.870	0.90	0.75	0.82
	(0.938)	(0.826)	(0.82)	(0.67)	(0.74)
ADA	0.959	0.874	0.91	0.76	0.83
	(0.932)	(0.840)	(0.75)	(0.72)	(0.73)
XGBoost	0.962	0.883	0.92	0.78	0.84
	(0.943)	(0.823)	(0.86)	(0.67)	(0.76)
MLP	0.932	0.824	0.78	0.68	0.72
	(0.932)	(0.822)	(0.78)	(0.67)	(0.72)
KNN	0.909	0.812	0.65	0.68	0.66
	(0.900)	(0.812)	(0.60)	(0.69)	(0.64)
SVM	0.817	0.787	0.39	0.75	0.52
	(0.871)	(0.802)	(0.51)	(0.71)	(0.59)

of Figure 5.3. The number of observations greater than 2 magnitudes brighter than the median of a rolling window has by far the greatest influence in discriminating between the classes.

5.4.1.2 Light curve only model

The best-performing binary task light curve-only model was XGBoost (CV F1-score of 76%). The implementation was performed with 150 Decision Trees at a learning rate of 0.2 and a maximum tree depth of 6. The Random Forest models follow closely behind (CV F1-score of 74%); the use of McNemar's test again shows XGBoost makes errors in the same proportions ($0.766 \leq p \leq 0.88$). The CV F1-score performance for this XGBoost model drops compared to the full feature model by 8 percentage points due to an increase in the number of false negatives from 69 to 93 and an increase in the number

TABLE 5.6: 4 class classification scores. Score with and without brackets, and abbreviations are as described in Table 5.5.

Model	Accuracy	Balanced Accuracy	CV Precision	CV Recall	CV F1-score
RF100	0.964	0.941	0.92	0.85	0.88
	(0.922)	(0.835)	(0.80)	(0.79)	(0.80)
RF250	0.964	0.936	0.92	0.85	0.88
	(0.924)	(0.835)	(0.81)	(0.79)	(0.80)
RF500	0.965	0.941	0.92	0.85	0.88
	(0.923)	(0.830)	(0.80)	(0.80)	(0.80)
RF750	0.965	0.942	0.92	0.86	0.89
	(0.923)	(0.833)	(0.80)	(0.80)	(0.80)
RF1000	0.965	0.942	0.92	0.86	0.89
	(0.923)	(0.833)	(0.80)	(0.80)	(0.80)
ADA	0.959	0.925	0.90	0.82	0.86
	(0.897)	(0.798)	(0.75)	(0.75)	(0.75)
XGBoost	0.962	0.928	0.91	0.84	0.87
	(0.922)	(0.820)	(0.83)	(0.77)	(0.80)
MLP	0.926	0.873	0.80	0.76	0.78
	(0.910)	(0.793)	(0.73)	(0.71)	(0.76)
KNN	0.895	0.795	0.90	0.48	0.63
	(0.898)	(0.730)	(0.86)	(0.67)	(0.75)
SVM	0.891	0.798	0.62	0.70	0.66
	(0.874)	(0.758)	(0.76)	(0.61)	(0.68)

of false positives to 36 from 20. Out of the 307 test set CVs, 214 were correctly identified (see bottom panel of Figure 5.1). The model AUC score also drops from 0.975 to 0.9622 (bottom panel of Figure 5.2). The number of observations greater than 2 magnitudes brighter than the median of a rolling window remains the feature that has by far the greatest influence in discriminating between the classes (bottom panel of Figure 5.3).

5.4.2 4 class classification

5.4.2.1 Full feature model

A 750-tree Random Forest model performs equally well or better than its competitors in each of the performance metrics evaluated for this 4-class full-feature task. The F1-score for CV classification stands at 89% though the remaining ensemble learning models follow closely behind. The model was trained such that only 25% of features (selected at random) could be used within each tree, with a maximum tree depth of 25. The confusion matrix (top panel of Figure 5.4) displays a strong performance in distinguishing CVs

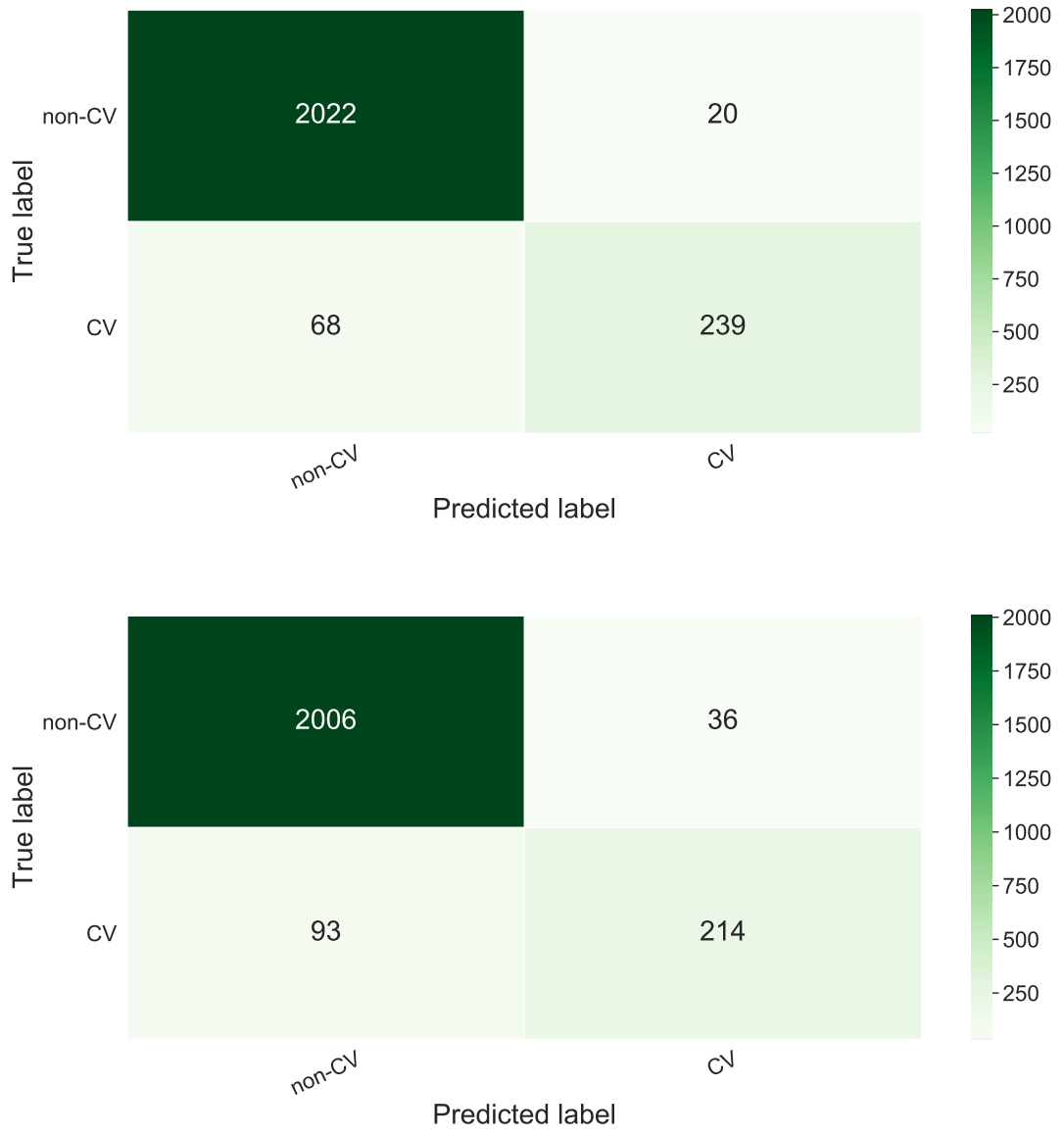


FIGURE 5.1: Confusion Matrices (CM) for the best performing binary task full feature (top) and light curve only (bottom) models. In each case, this was an XGBoost model — achieved the highest F1-score. The CMs show the numbers corresponding to precision, recall and accuracy scores in Table 5.5. There are over 6 and a half times more non-CVs in the test set than CVs, raising the overall accuracy score, the balanced accuracy score is more able to account for this class imbalance.

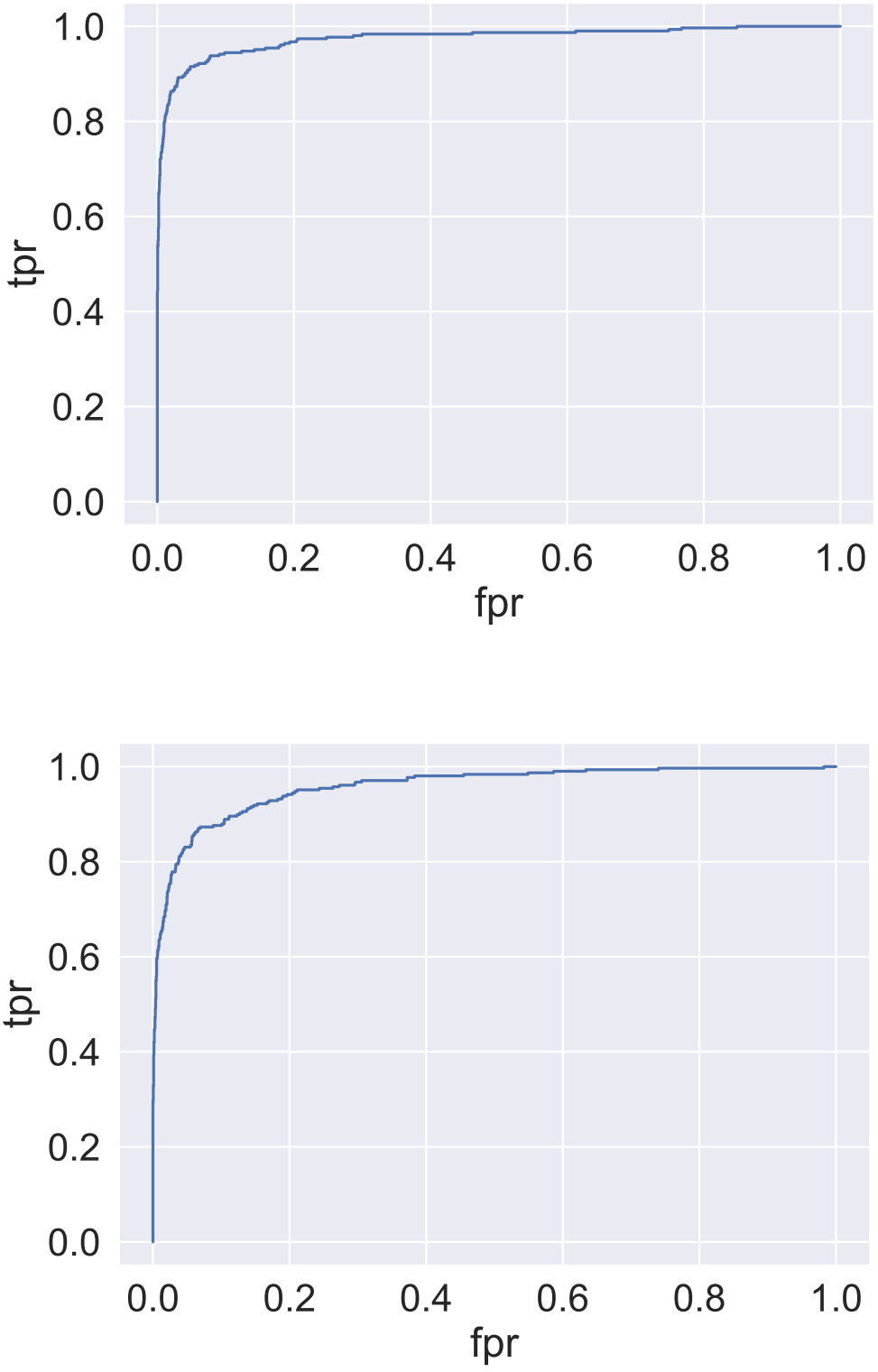


FIGURE 5.2: ROC curves for the full feature and light curve only binary task models achieving the highest CV F1-scores. On the top is the curve for the full feature model, while on the bottom is that for the light curve-only model. The full feature model area under the curve is 0.975, for the light curve-only model this is 0.9622, indicating a strong performance in each case.

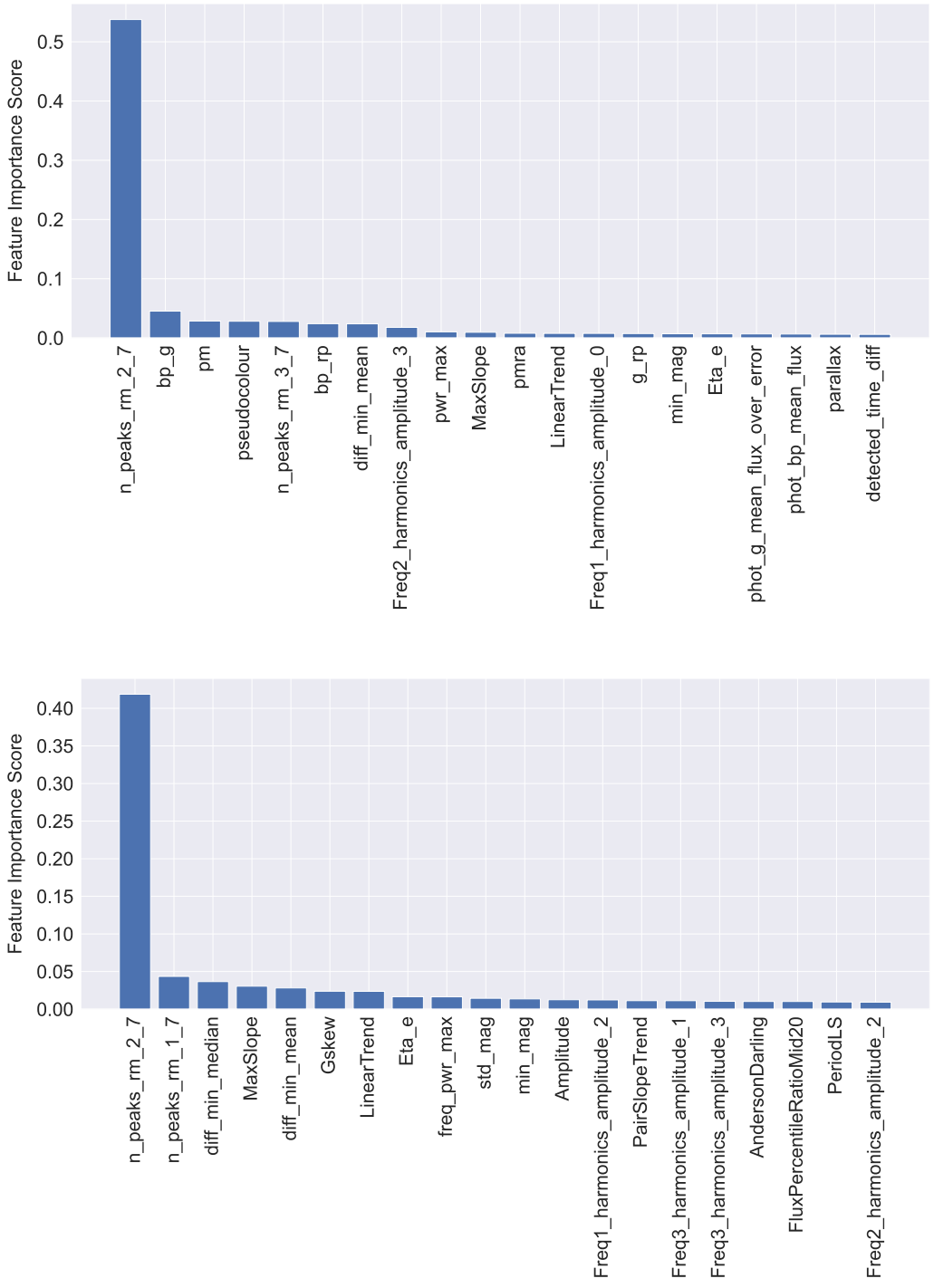


FIGURE 5.3: Feature importance scores for the 20 most influential features within the best performing full-feature and light curve only binary task models. Feature importance refers to a class of techniques for assigning scores to input features to a predictive model, in this case, XGBoost, that indicates the relative importance of each feature when making a prediction. The most important feature for each of the full feature (top) and light curve only (bottom) models is *n_peaks_rm_2_7* — number of instances of data points at least 2 magnitudes brighter than the median of a rolling window of 7 epochs. Feature definitions are contained in Tables 5.1, 5.2 and 5.3.

from other classes. Those CVs that were misclassified were mostly predicted to be of the SNe class (39/44). The top panel of Figure 5.5 presents histograms of the probabilities of class assignment for this model. The vast majority of test set examples, 274 out of the 285 predicted CVs, were predicted as such with probabilities greater than 50%. 79 of the 285 were predicted as CVs with a probability of 95% or above. All but 3 examples predicted as YSOs are classified with 50% probability or higher.

According to the feature importances (top panel of Figure 5.6) the temporal baseline of observations (*detected_time_diff*) has the greatest influence in discriminating between classes. In addition to Gaia's observing strategy and their prevalence in the dataset, this can be partially explained by the properties of the majority class, supernova — they are too distant for their progenitors to be observable by Gaia, and after several months they become too faint to be observable above the light from their host galaxy. Of the supplementary features, parallax and proper motion are expected to provide the greatest ability in class distinction, with the ability to distinguish extragalactic sources from those nearby. They both appear high in feature importance, as do the right ascension and declination error features. These errors are noticeably higher for SNe (∼12.8 mas) than for remaining classes (∼0.08-0.17 mas) attributed to the ability to measure these properties being affected by crowding (including contamination of light from the host galaxy).

5.4.2.2 Light curve only model

A 1000 tree Random Forest model performed the best, achieving the highest CV F1-score (81%) for the 4 class light curve feature-only task, though the remaining ensemble learning models follow closely behind. The model was implemented with a maximum tree depth of 30 and 75% of randomly selected features available for each tree. While 247 of the 306 CVs have been correctly classified (bottom panel of Figure 5.4), the contamination of other classes into those targets predicted as CV increases from 8 to 18% compared to the full feature model. Like the full feature model, misclassified CVs are mostly assigned the SNe label. The histograms of class assignment probability (bottom panel of Figure 5.5) show the majority of CVs are predicted as such with greater than 50% probability, though more examples are now present in the tail of the distribution.

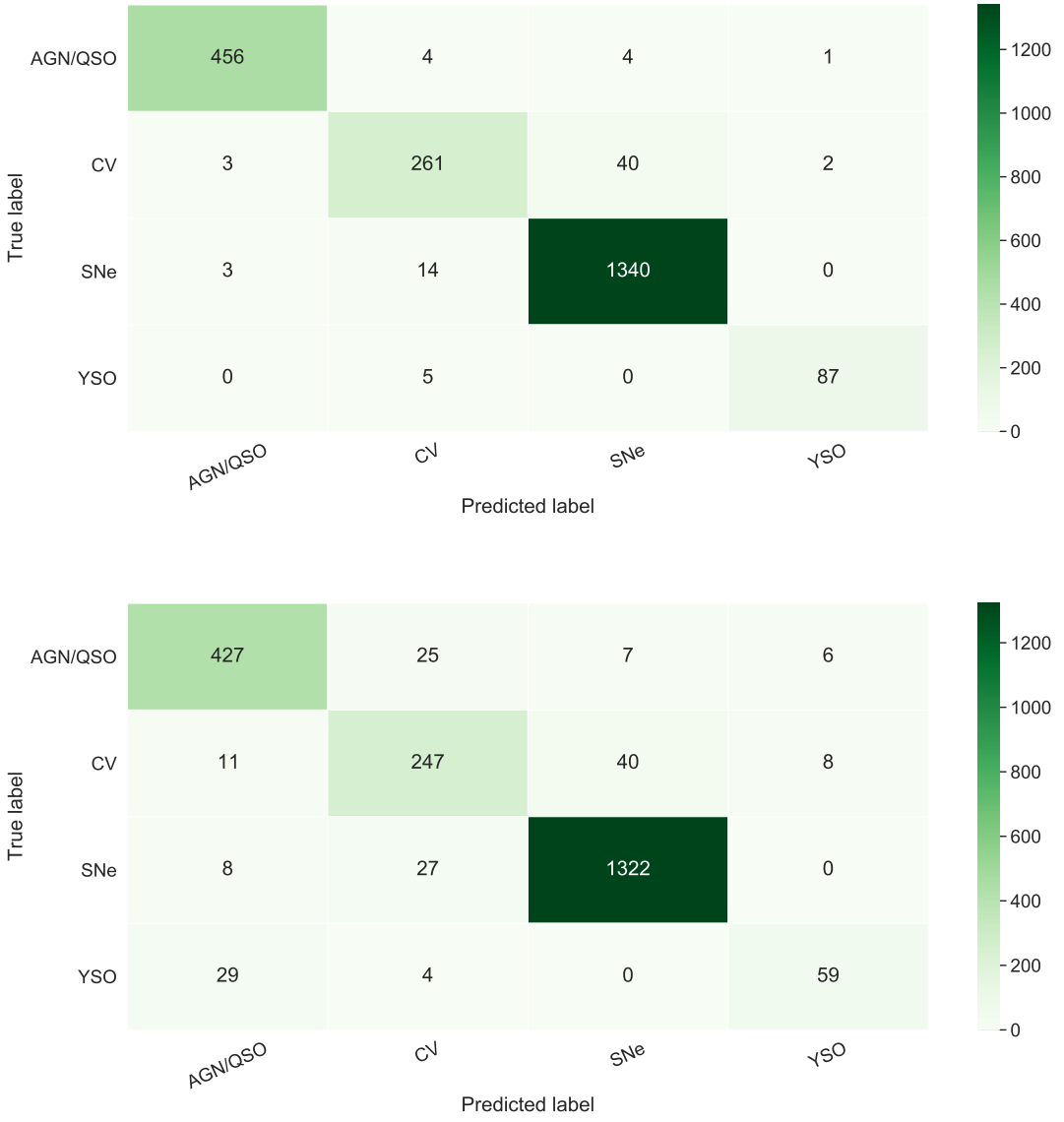


FIGURE 5.4: Confusion Matrices for the best performing full feature and light curve only models in the 4 class classification task. On the top is the 750-tree Random Forest model trained with the full complement of features. 262 of the 306 CVs in the test set were successfully classified (true positives), the majority of those misclassified, 39 of 44, were predicted to be supernovae. On the bottom is the 1000-tree Random Forest model trained with light curve-derived features only. Less true positives (247) compared to the full feature model. Also an increase in the number of false positives from 23 to 56, of which the majority were AGN and supernovae.

According to the feature importances (bottom panel of Figure 5.6) the temporal baseline of observations (*detected_time_diff*) also has the greatest influence in class distinction.

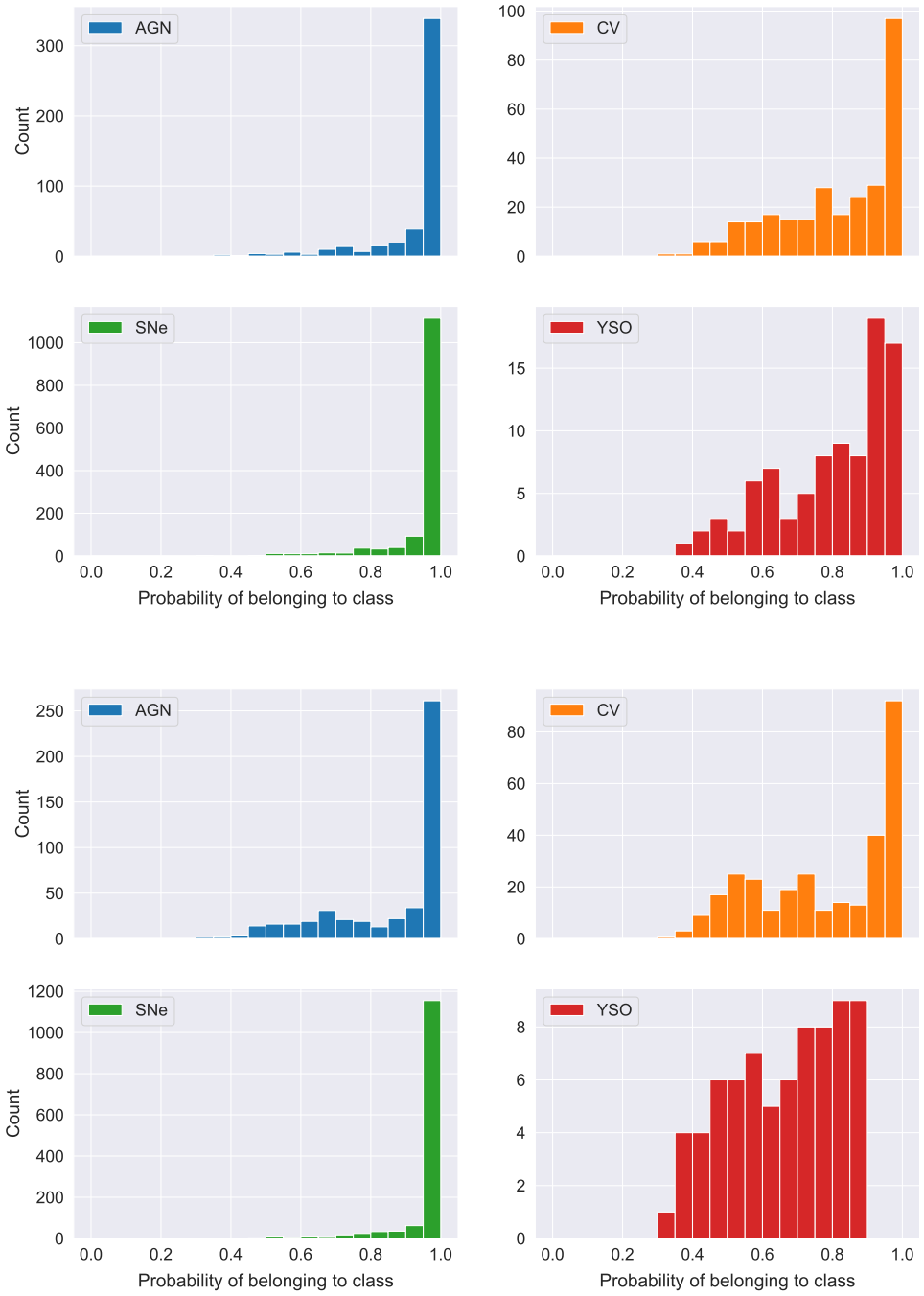


FIGURE 5.5: The number of test set examples predicted as CV (orange), AGN (blue), SNe (green), and YSO (red) separated in bins of probability of class association calculated for the full feature and light curve only 4 class models with the highest F1-scores. Each tree in the Random Forest model predicts class probabilities for each example — these are the fraction of samples of the same class in the associated leaf evaluated during training. These probabilities are averaged for the forest prediction. Class probabilities for the full feature Random Forest model (top) show nearly all examples are assigned classes with greater than 50% probability, the majority of which are in the 95-100% bins. For the light curve only features 4 class model (bottom), one can say likewise, however, the YSO class assignment probabilities are more uncertain.

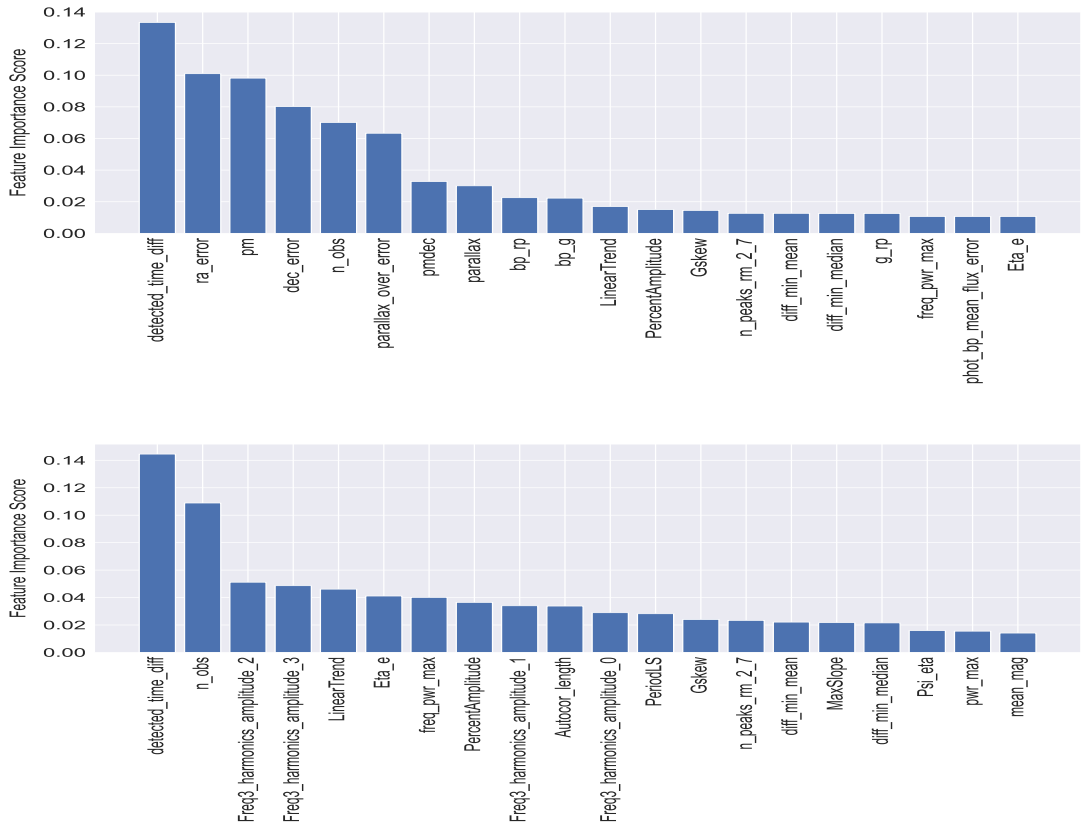


FIGURE 5.6: The top 20 features based on feature importance scores for the 4 class full feature and light curve only models with the highest F1-scores. The full feature model's best-performing feature (top) was the time between the first and last observation of the target, followed by the error in the right ascension, proper motion, and the error in declination. The same best-performing feature is present for the light curve-only model (bottom). Feature definitions are contained in Tables 5.1, 5.2 and 5.3.

5.5 Discussion

5.5.1 Semi-regular, short duration outbursts

The light curve feature that logs the number of epochs that are at least 2 magnitudes brighter than the median of a rolling window of 7 epochs, *n_peak_rm_2_7*, outperforms all others in feature importance for the best performing full feature and light curve only binary models. The semi-regular, short-duration outbursts of DNe are effectively picked out using this feature as found during its development. Such characteristics are more likely to be identified within Gaia's transient alerts pipeline than the less frequent alert-triggering features of other CV subtypes so the high ranking of the feature may be expected. Indeed, a coordinate cross match with 'The Catalogue and Atlas of

Cataclysmic Variables'[6] (Downes & Shara, 1993; Downes et al., 1997), reveal 77% of successfully cross-matched dataset CVs are listed as being DNe. Exploration of the true positives for each of the best-performing binary models reveals the majority display the expected dwarf novae morphology (78 and 73% for the full feature and light curve-only models respectively).

5.5.2 Limited Epoch Photometry

A significant fraction of the dataset is constructed from target light curves with few epochs of observation, 36% of targets contain 5 or fewer datapoints in their light curves. This is due to the combination of Gaia's sampling frequency and systems too faint to be observed by Gaia until a brightening event propels them into visibility. Transient phenomena more likely to display this trait will be those exhibiting a rapid and large amplitude brightening, for example, SNe and the CV subclasses of classical and dwarf novae. Considering SNe comprise the majority (58%) of the dataset, this may explain the strong performance of *detected_time_diff* (temporal baseline of observations) in class distinction (see Figure 5.6). It may also explain the difficulty that the best-performing 4-class models have in distinguishing CVs from SNe. Of the CVs misclassified by the best performing full feature 4 class model, 87% (39 of 45) are predicted to belong to the SN class, while for the corresponding light curve only model, 68% (40 of 59) are predicted as SNe. Similarly, the majority of misclassified SNe in each of those models are predicted to be of the CV class. Inspection of the CVs misclassified as SNe reveals the majority possess light curve morphologies that are present for the SNe samples — those with few data points (2–10 observations) and those exhibiting an approximately exponential decline with no pre-explosion data.

5.5.3 Metadata and high imputation

A McNemar's test suggests the use of metadata has an impact on model performance when comparing the full feature and light curve only XGBoost models ($p = 10^{-7}$). However, the small difference in classification accuracy between these two binary models (1.9%) indicates that the addition of survey metadata provides minimal benefit in distinguishing CVs from non-CVs. This is also shown by the small difference in the

[6] `https://archive.stsci.edu/prepds/cvcat/index.html`

AUC (1.3%) between these models, with both performing strongly by this measure (0.975 and 0.962 for the full feature and light curve only models respectively). The feature importances for both binary models further illustrate this point — the influence of supplementary features in class distinction is dwarfed by the light curve derived *n_peaks_rm_2_7* feature. Either the metadata is unimportant or mean imputation has diluted the influence this data has on class distinction. The latter seems more likely when presented with pair plots of Figure 5.7 that show transient classes in metadata feature space. This plot is of particular use in interpreting the performance of algorithms that rely on class separation within feature space (e.g., KNN and SVM). Evident is the distinction between YSOs from CVs, SNe, and AGN in colour space (bp-rp, bp-g, g-rp); and CVs and YSOs from SNe and AGN when proper motion is considered.

The use of mean imputation has its drawbacks, it ignores relationships between features, the correlation for example, and reduces the variance of the variable, thereby introducing bias to the model. Furthermore, the strategy may not be suitable for several supplementary features. For example, the parallax may not be measurable because the object is too far away (too small to measure); and a missing value for proper motion can either be due to the object having no proper motion to measure or be due to it being too distant to be measured. A more appropriate strategy could be to replace these with a value of zero — a more accurate quantity for the parallax and proper motion of the most distant sources — though this does not account for the unavailability of these features due to an unsuccessful cross-match with EDR3. While alternative methods of handling missing data could be employed (such as those summarised in Soley-Bori 2013), a large amount of data is missing for the supplementary features (58-90%), this can limit the effectiveness of any such strategy (Jäger et al., 2021). Figure 5.7 shows how photometric colour information can be an important property for class distinction, this is readily available in multi-band surveys such as ZTF and can be used to help alleviate the issue.

5.5.4 Comparison with other work

The results of this investigation compare favourably with similar classification attempts where CVs are included as a class. Neira et al. (2020) experiments with CRTS light curves in their 8 class classification model yielded an F1-score of 75% for the CV class, while this work exceeds this in both the binary (76%) and 4 class (80%) tasks where only

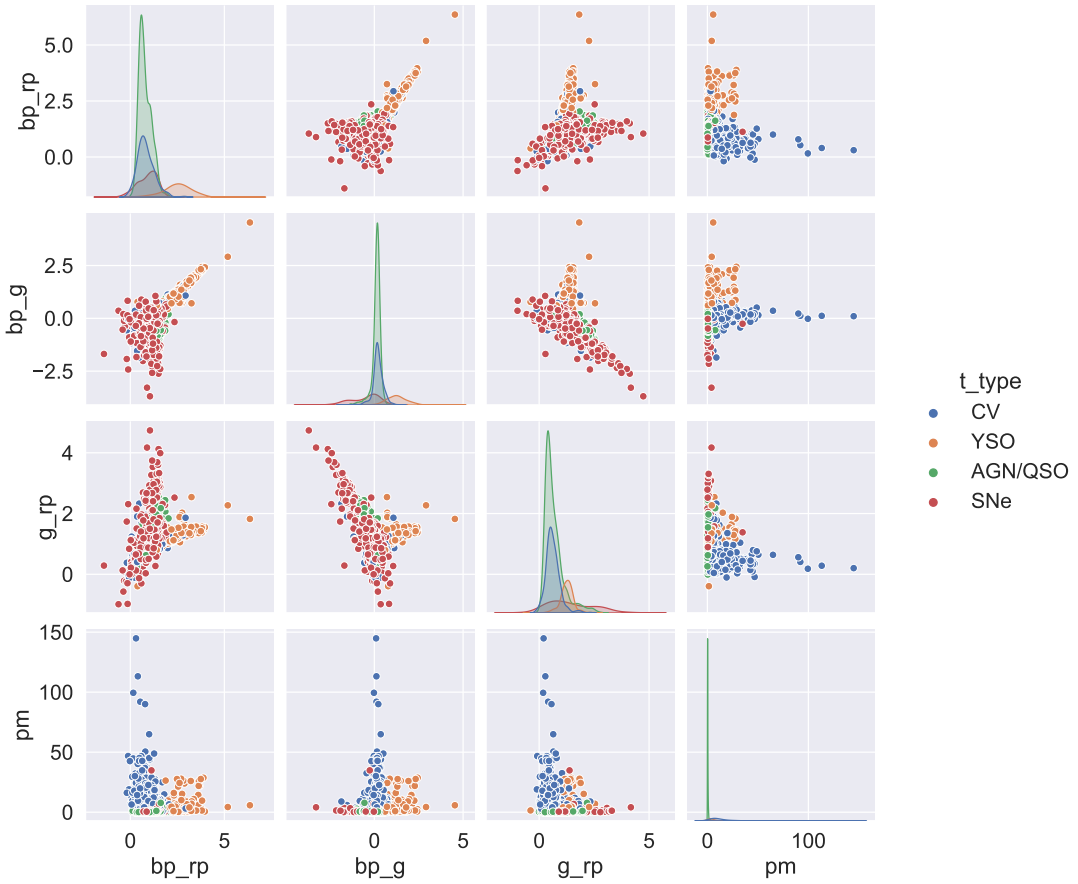


FIGURE 5.7: The pairplot allows us to see both the distribution of the single variables (plots shown diagonally from top left to bottom right) and relationships between two variables (off-diagonal plots). This is shown for the bp-rp, bp-g, and g-rp colours, and proper motion. YSOs are redder in colour compared to SNe, CVs, and AGN, observed in both the single variable distribution and relationship plots, thus allowing for a significant level of class separation. Introduction of proper motion allows for the separation of the more distant (extragalactic) SNe and AGN from the closer (Galactic) CV and YSO population.

light curve features are used. Sánchez-Sáez et al. (2021) evaluated 3 different algorithms in their tiered classification attempts to distinguish between CVs, SNe subclasses, AGN, YSOs and variable star subclasses from a dataset constructed from ZTF light curves and colours from ALLWISE. Their CV recall scores for their implementation of the Balanced Random Forest (Chen & Breiman, 2004), XGBoost, and Multi-Layer Perceptron classifiers are 68%, 72%, and 61% respectively. This compares with 67% and 80% for my light curve only best-performing binary and 4 class models respectively. These comparisons do not however take into account differences in the instruments used to collect the data, which translates to the nature of photometric data (e.g., observing cadence, waveband). Furthermore, comparisons do not consider differences in transient classes to classify and

ML methods employed.

5.5.5 Gaia Unknowns

5.5.5.1 Model predictions on unknown sample

The model that produced the highest CV F1-score overall — full feature 4 class model (Random Forest with 750 trees) — is used to make class predictions of targets labelled as 'unknown' (unclassified) within the Gaia alerts stream. As of December 2021, 13,241 targets were of 'unknown' class. Of these, the model predicted 2,833 (21%) to be of the CV class, 1,928, 6,611, and 1,869 were classified as AGN/QSO, SNe, and YSOs, respectively. As mentioned in Section 5.2, the unknown sample will contain several minority classes (e.g., microlensing and tidal disruption events) not included in the test set used to evaluate model performance. I aim to assess the impact this has on my model's ability to generalise to the unknown sample and new transient alerts in general. This will require spectroscopic observations for a sufficient number of the 2,833 predicted CVs to identify their true transient classification.

5.5.5.2 Spectroscopic follow-up

I am therefore undertaking a pilot study to assess the performance of the model and the methods used by obtaining spectroscopic observations to classify those targets that can be observed with the SPRAT low-resolution spectrograph (Piascik et al., 2014) mounted on the Liverpool Telescope (LT; Steele et al. 2004). These spectra cover a wavelength range of 4000 to 8000Å with a resolution of 18Å, corresponding to a resolving power, R=350, at the centre of this range. A limit on telescope time and the need for high-quality spectra requires an efficient observation strategy. Accordingly, observations are limited to targets with a median brightness no fainter than 18th magnitude. Furthermore, only those targets that rise highest in the sky — visible for longer at a lower airmass — are considered. Therefore, the sample is limited to those with a declination corresponding to an altitude no lower than 50 degrees when at transit altitude. These cuts leave a sample of 220 targets, 7.8% of the total catalogue — a representative fraction with which one can validate the performance of the model. I have spectroscopically classified 15 of this sample, all of which I can confirm are of the CV class. Details of

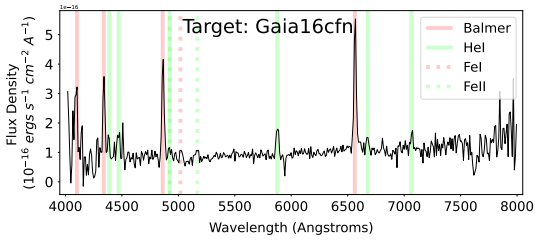
Target	**Classification**	**Comment**
Gaia16cfn	CV (Dwarf Nova)	Clear Balmer and He I emission. He I λ4922 blended with Fe II λ4924. Characteristic of DNe subtype.
Gaia17ccv	CV (Decline from dwarf nova outburst)	CV on decline from outburst, Faint Hα and Hβ emission, He II λ4686 in emission. Double peaked lines — indicative of high inclination system.
Gaia17dfn	CV	Balmer and He I λ4471 lines in emission
Gaia18auz	CV	Clear Balmer emission with several faint He I lines in emission
Gaia18dgt	CV (Dwarf Nova)	Broad Balmer emission with lines of He I. He I λ4922 blended with Fe II λ4924. Characteristic of DNe subtype. Double peaked emission, possible high inclination system
Gaia18dhv	CV	Balmer, He I and He II in emission
Gaia19bzn	CV	Clear Balmer emission; faint lines of He I and He II
Gaia19cln	CV	Clear Balmer emission; Lines of He I and He II also present; He I λ4922 blended with Fe II λ4924
Gaia20air	CV	Clear Balmer emission; Lines of He I and He II also present; He I λ4922 blended with Fe II λ4924
Gaia20bjd	CV	Clear Balmer emission; Lines of He I also present; He I λ4922 blended with Fe II λ4924
Gaia20cpq	CV (Dwarf Nova)	Clear Balmer emission; Lines of He I and He II also present; He I λ4922 blended with Fe II
Gaia21beh	CV	Outburst spectrum. Possible very faint Hα absorption, clear absorption in remaining Balmer lines and He I λ4471, He II λ4686 in emission (faint).
Gaia21cgv	CV	Balmer and He I emission lines, faint Fe II λ5169
Gaia21cul	CV	Clear Balmer and He I emission lines
Gaia21eyb	CV	Balmer, He I, He II and Fe II emission lines, He I λ4922 blended with Fe II λ4924

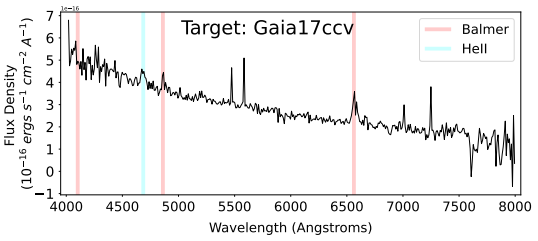
TABLE 5.7: Classifications based on LT SPRAT spectroscopy of several targets labelled as 'unknown' (without a transient class assignment) within Gaia Science Alerts and predicted as CV by the RF750 model.

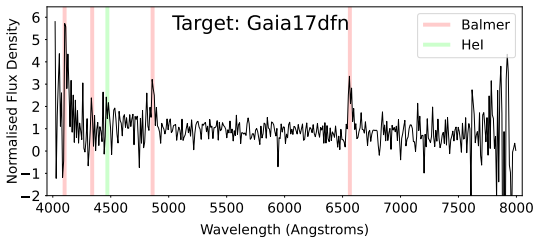
these targets are given in Table 5.7, while the associated SPRAT spectra are shown in Figure 5.8. Classification as a CV is based on the presence of Balmer and/or He I/He II lines. Where the signal-to-noise ratio of the spectrum permits, subtype classification is performed. Full details of the spectral features used for classification are given in Szkody (1998) and Hou et al. (2020).

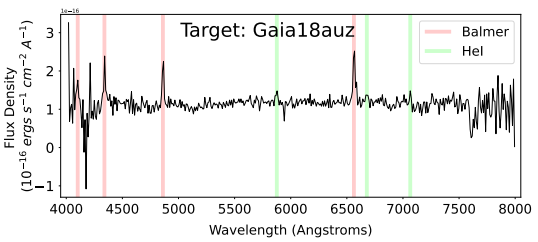
5.6 Conclusions and future work

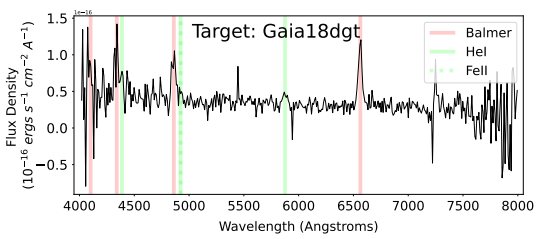
The advent of wide-field synoptic surveys has revolutionised time-domain astronomy with their ability to detect millions of transient events per night. The use of Machine

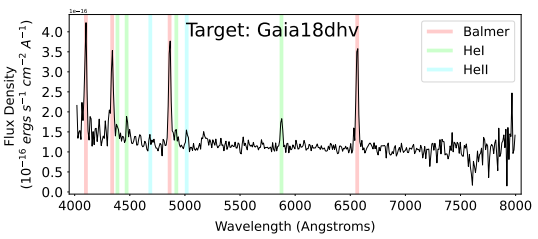

Target: Gaia16cfn
Balmer
HeI
FeI
FeII
Flux Density
Wavelength (Angstroms)

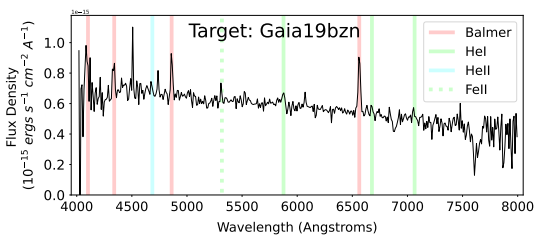

Target: Gaia17ccv
Balmer
HeII
Flux Density
Wavelength (Angstroms)

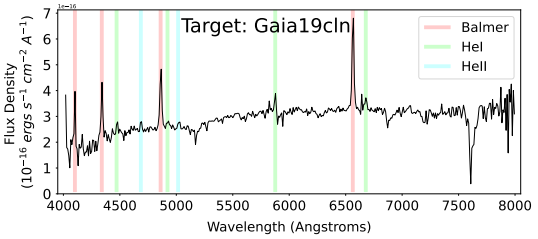

Target: Gaia17dfn
Balmer
HeI
Normalised Flux Density
Wavelength (Angstroms)

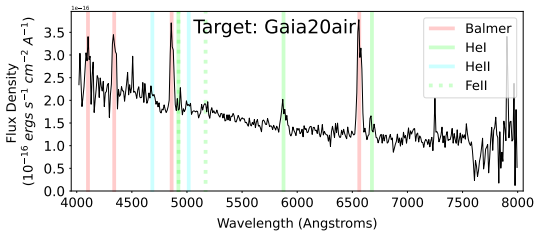

Target: Gaia18auz
Balmer
HeI
Flux Density
Wavelength (Angstroms)

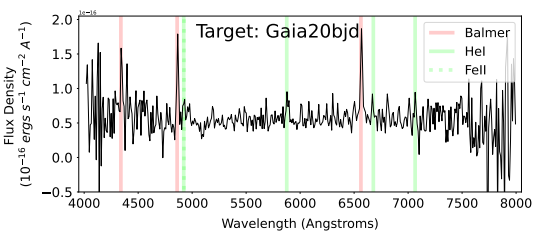

Target: Gaia18dgt
Balmer
HeI
FeII
Flux Density
Wavelength (Angstroms)

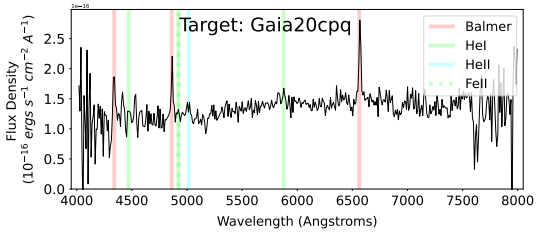

Target: Gaia18dhv
Balmer
HeI
HeII
Flux Density
Wavelength (Angstroms)

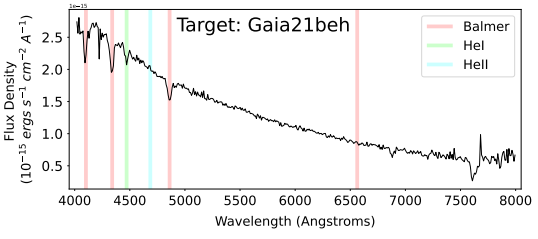

Target: Gaia19bzn
Balmer
HeI
HeII
FeII
Flux Density
Wavelength (Angstroms)


Target: Gaia19cln
Balmer
HeI
HeII
Flux Density
Wavelength (Angstroms)


Target: Gaia20air
Balmer
HeI
HeII
FeII
Flux Density
Wavelength (Angstroms)


Target: Gaia20bjd
Balmer
HeI
FeII
Flux Density
Wavelength (Angstroms)


Target: Gaia20cpq
Balmer
HeI
HeII
FeII
Flux Density
Wavelength (Angstroms)


Target: Gaia21beh
Balmer
HeI
HeII
Flux Density
Wavelength (Angstroms)

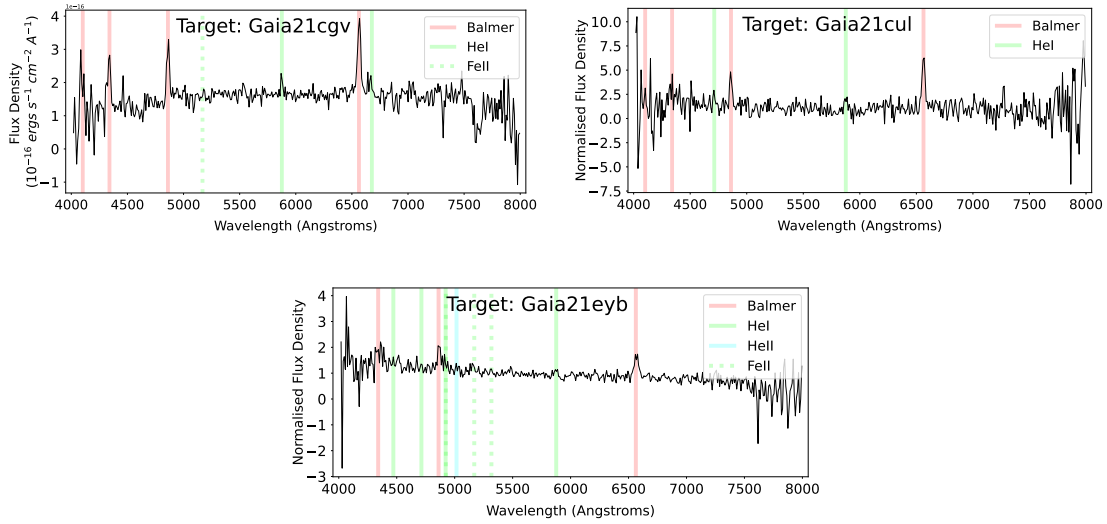


FIGURE 5.8: SPRAT spectra of targets in Table 5.7. Spectral lines are indicated in plots, labelled in the legend for each.

Learning is recognised as the best method of source classification for this deluge of transient sources. Machine Learning algorithms have been applied widely to data from several surveys including CRTS and ZTF photometry. In this work, I applied ML techniques to the transient stream of Gaia Science Alerts, a resource not fully explored with ML. My focus lies in the identification of Cataclysmic Variable stars, a class of transients providing ideal laboratories for the study of accretion and binary evolution. Using features extracted from light curves of classified sources and associated metadata as input, I evaluated the use of Random Forest, AdaBoost, XGBoost, K Nearest Neighbours, Support Vector Machines, and a Multi-Layer Perceptron in performing several tasks. These are the identification of CVs in the context of binary classification (CV or non-CV) and a 4-class task (CV, AGN, SNe, YSOs). Each of these tasks was performed with and without metadata (e.g., Gaia parallaxes and colour) during training. By comparison of the F1-score of all models across both tasks, the 4 class Random Forest model trained with both light curve and metadata-based features performed the best with an F1-score of 89% when evaluated on the test set. I applied this model to the list of unclassified targets within GSA. The model predicted 2,833 of these 'unknowns' to be of the CV class. I am now undertaking a spectroscopic observing campaign to spectroscopically classify a representative fraction of these targets to validate the model's performance. So far, I have been able to spectroscopically confirm 15 targets to be of the CV class.

The use of data beyond light curve features seems necessary in order to achieve classification performance close to an F1-score of 90%. However, with light curve features alone the performance of the model compared well with other works, despite more sparsely sampled light curves. The lessons learnt during this exploration of the GSA resource and the classified targets from my spectroscopic database of targets will be useful in the next phase of research. This will be the application of ML to the multiband high-cadence light curves of the ZTF survey.

The next phase will be an opportunity to explore methods of handling class imbalance and missing data. Class imbalance, present within the dataset (see subsection 5.2.1), tends to bias classifiers to recognise the oversampled class more than the undersampled class. Algorithm-specific solutions exist, for example, within Random Forest one may grow each tree with the same number of targets per class by oversampling or undersampling using the bootstrap sampling process (Fernández et al., 2018). Data augmentation methods (e.g., Wen et al. 2020) to generate new examples based on existing examples will also be explored. The use of mean imputation for handling missing data is simple and parameter-free. Whilst this method can cause biases (see subsection 5.5.3), I deemed the exploration of several imputation methods beyond the scope of this work, though it is something to be explored in work with ZTF data. The reliability of class labels will also be important for the next research phase and once LSST becomes operational. Whilst there is confidence in the methods employed in the labelling of examples used here (see subsection 5.2.1), I acknowledge that labelling errors do occur. This can add noise to the dataset, deteriorating classifier performance (Frenay & Verleysen, 2014) and reducing the effectiveness of performance optimisation techniques such as hyperparameter tuning.

The methods employed in this work are transferable to the data available from the ZTF survey. This data should provide the necessary information to identify subclasses within the CV population and pick out rare varieties that further our understanding of binary evolution. For example, the ∼2 day cadence provides the sampling necessary to recognise the defining characteristics of DNe subtypes, such as the superoutbursts of SU UMa systems (e.g., Szegedi et al. 2022) and the standstills of Z Cam systems (Simonsen et al., 2014); and identify characteristics present in outbursting AM CVns, such as the short duration rebrightenings on the fading tail of a superoutburst (Kato & Kojiguchi, 2021). The ability to automatically distinguish between the different CV subtypes will depend upon several factors, one of which is the quality of features. Several

features used in this work have so far shown their effectiveness at class distinction, others may become more significant once computed with the higher cadence data, while the development of features geared towards the identification of specific subtypes should provide further benefit. The prevalence of a given subtype within the dataset is another factor that I expect to impact classifier performance. The sensitivity of a survey to certain CV subtypes results in the under-representation of novae, AM CVns and nova-likes compared to DNe due to the rarity of eruptions, faintness, and photometric stability, respectively. This is where the methods of handling class imbalance described in the previous paragraph will become invaluable. The methods used here and the lessons learnt will aid in the goal of separating the rare CV systems from those more common and hopefully lead to a greater understanding of binary evolution.

Chapter 6

ZTF Machine Learning Applications

6.1 Introduction

A specific focus on the automated identification of CVs and their subtypes is an active yet underdeveloped field of research. Examples to date include: the identification of 497 CVs from ZTF alerts using simple colour, amplitude, and variability timescale filters (Szkody et al., 2020, 2021); an extension of this filtering approach by van Roestel et al. (2021), which utilised Gaia and Pan-STARRS colours to identify nine outbursting AM CVns within ZTF alerts; and the application of machine learning to identify CVs in Gaia Science Alerts (Mistry et al., 2022).

Here are presented details of the development and application of an automated ML pipeline to identify the various classes of CVs from the ZTF alert stream via the Lasair alerts broker (Smith et al., 2019). I start by explaining the initial alerts filtering using Lasair (Section 6.2.1) before moving on to describing the construction of the dataset upon which an ML classifier is generated (Sections 6.2.2 – 6.2.5). Sections 6.2.6 – 6.2.10 describe the ML techniques adopted and algorithms tested. The results of my efforts to generate a suitable ML CV classifier for my pipeline are presented in Section 6.3 along with its initial outcomes based on implementation. The discussion of my results (Section 6.4) will be given in the context of light curve profiles and the underlying physical properties of the CV subtypes.

6.2 Method

6.2.1 Alerts filter

The alert stream from ZTF is ingested by alert brokers such as Lasair (Smith et al., 2019) and Alerce (Förster et al., 2021). They provide real-time alert access, science, difference and reference image cutouts, light curves of the associated ZTF object, contextual information, statistics derived from source photometry, and the ability to cross-match events with catalogued sources. Brokers provide the ability to filter alerts based on the above to focus on those that are most relevant to their science goals. My pipeline experiments with Lasair's cross-matching and filtering services to focus on objects within the typical parameter space of CVs as a first stage before implementing my ML classifier.

To remove non-CV catalogued sources, the Sherlock classification software (Smith et al., 2020), implemented by Lasair for cross-matching, is examined. Sherlock uses a model, generated by a boosted decision tree algorithm, that mines a database of historical and ongoing astronomical survey data to predict the nature of the object based on the resulting crossmatches. The database includes datasets from all-sky surveys as well as more source-specific catalogues such as the Million Quasars Catalog (Flesch, 2019), Downes Catalog of CVs (Downes et al., 2001), and the Ritter Cataclysmic Binaries Catalog v7.24 (Ritter & Kolb, 2003). Sherlock assigns the label Variable Star (VS), Cataclysmic Variable (CV), Active Galactic Nuclei (AGN), or nuclear transient (NT) should the transient be located within the synonym radius (1.5") of a catalogued point source or, in the case of a NT, the core of a resolved galaxy; a supernova (SN) if not classified as a NT but is found close enough to a resolved galaxy to be deemed physically associated; a Bright Star (BS) if the transient is not matched against the synonym radius of a star but is associated within the magnitude-dependent association radius; Orphan if the transient fails to be matched with a catalogue source; or Unclear otherwise. To limit alerts of non-CVs, I made use of Sherlock and catalogue cross-matching in the manner described in Section 6.3.4.

The remaining sources are subject to colour and magnitude change cuts akin to those described in Szkody et al. (2020, 2021). In those works, the ZTF alert stream filtering involved looking for point sources with g-r colour < 0.6 and a magnitude change $\Delta m >= 2$ within a timescale of 2 days in the g band. This resulted in a total of 701 known or

candidate CVs over two years of its implementation that typically displayed dwarf nova outbursts and changes in accretion state. I relaxed these constraints with respect to Szkody et al. (2020, 2021) to maximise the number of targets for classification. In performing a cut based on colour, attempts were made to account for several factors: differences in the sampling between the g and r band; sampling differences between outburst activity and quiescence; and the tendency of CVs to have bluer colours during outbursting phases than during quiescence (a consequence of the enhanced accretion and increased temperature of the disk during outburst). Therefore, for each source, the colour for each night of observation was extracted (where calculable); the mean and median averages of these were recorded along with the colour at maximum and minimum brightness. The constraint of $<= 0.7$ for each of these quantities, as well as for the overall mean colour (calculated without the epochal requirement) was utilised. Figure 6.1 shows that a significant fraction of CVs will be recovered at or below the epochal mean g-r of 0.7. This constraint is flexible, based on the type of CV I may wish to focus my attention on. Constraints placed on magnitude change, Δm, involved experimenting with various thresholds. A higher Δm yielded sources with more rapid variability, e.g., Z Cam systems, while lower values increased the contribution of sources akin to nova-likes. Given that alerting sources that the filter outputs are entered into an ML classifier to distinguish these variability differences, foregoing a Δm constraint is the approach adopted.

6.2.2 Source List

The light curves and associated metadata (see the following subsection) of the sources remaining after the Lasair filter are used as input for an ML-based CV subclass classifier. The classifier is trained on the ZTF g and r band light curves of catalogued CVs whose subtypes have been ascertained along with associated Gaia Data Release 3 data (Gaia-Collaboration et al., 2022) where available. This section describes the nature of the data set for training and testing of candidate classifiers.

To construct a dataset, I consulted the American Association of Variable Star Observers Variable Star Index (VSX[1]) which is a continuously updated repository of transient sources. Confirmed CVs from archival resources such as the Catalogue and Atlas of

[1]https://www.aavso.org/vsx/index.php

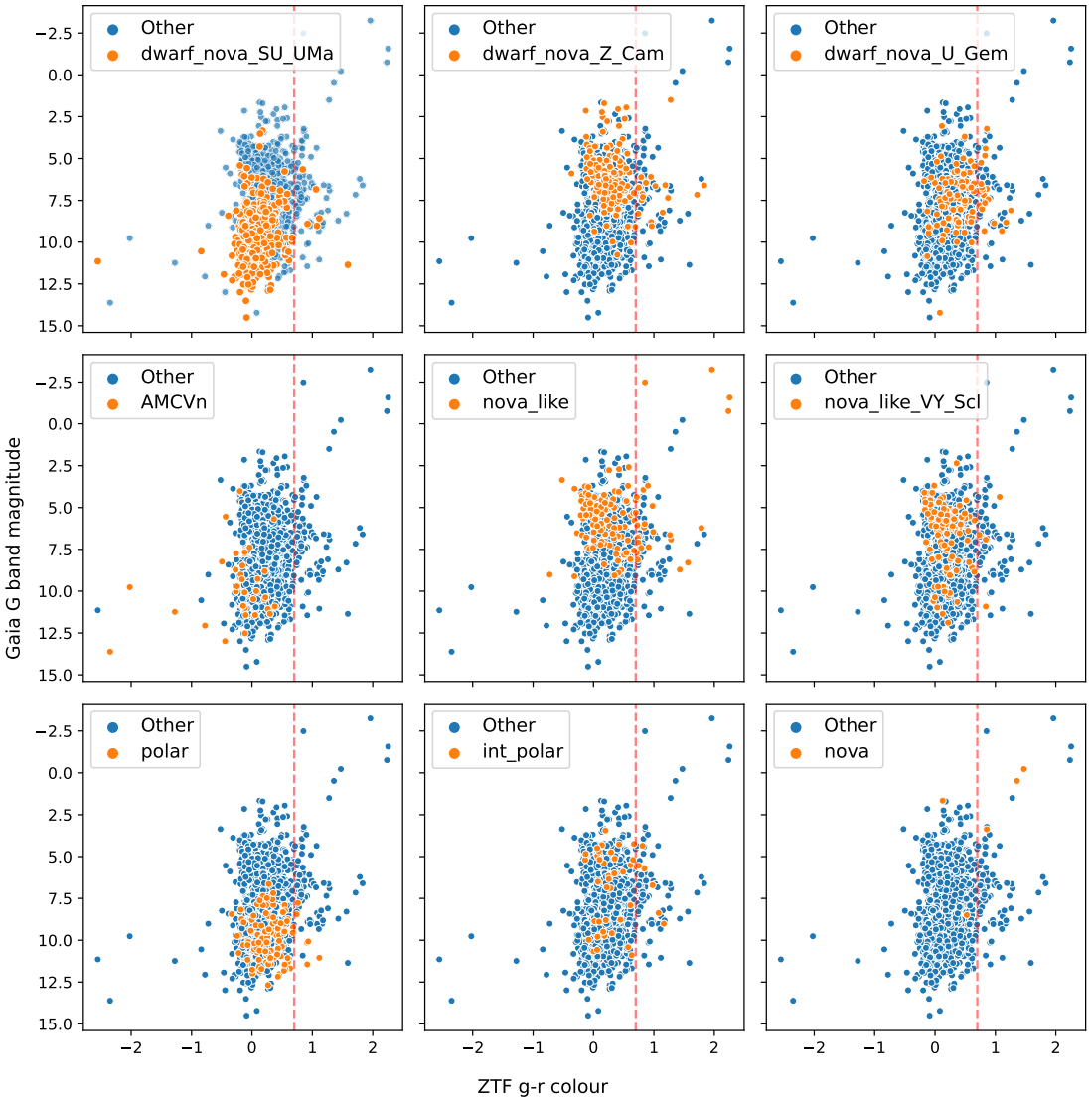


FIGURE 6.1: Colour magnitude diagrams using Gaia G band absolute magnitude and the colour derived from the ZTF g and r bands. The dashed red line in each plot denotes the ZTF g-r colour threshold of 0.7. Orange points in each subplot denote examples of a particular CV class, while the blue points represent examples belonging to the remaining classes (labelled 'other').

Cataclysmic Variables [2] (Downes et al., 2001), and the Catalogue of Cataclysmic Binaries, Low-Mass X-ray Binaries and Related Objects [3] (Ritter & Kolb, 2003) are contained within the repository, as are more recent discoveries detailed in literature (e.g., Wenger et al. 2000; Szkody et al. 2020; van Roestel et al. 2022). Each repository source has a dedicated page where further information can be found such as their designated names in other surveys, references to literature for that source, orbital periods, and more. The labelling procedure conducted by VSX involves constant review and revision of metadata,

[2] https://heasarc.gsfc.nasa.gov/W3Browse/all/cvcat.html
[3] https://heasarc.gsfc.nasa.gov/W3Browse/all/rittercv.html

with citations for any new details and rationales behind changes fully documented. VSX contained a list of over 15,300 targets classified as CV, of which 5,683 were successfully cross-matched with ZTF alerts objects. I supplemented this list with novae catalogued in the Bright Transient Survey [4] (BTS; Perley et al. 2020) and not in the AAVSO list. This constituted an extra 28 sources making a total of 5,708 CVs. The vast majority (4,822) were of the dwarf nova subclass. Since a more granular classification than this is the aim, the sample is refined further to only include dwarf nova examples with further subdivision into the U Gem, Z Cam, and SU UMa subtypes. This resulted in a dataset of 1,568 samples.

6.2.3 Light curves

The light curves themselves are generated from observations with the 47 square-degree camera mounted on the Samuel Oschin Telescope at Palomar Observatory in California (Harrington, 1952). For a 30-second exposure, the median 5σ limiting magnitude is 20.8 in the g band and 20.6 in the r band. The observing strategy involves three surveys, the g and r band data for two of which are available publicly. The Northern Sky survey is a three-day cadence survey of all fields north of declination $-31°$, while the Galactic plane survey observes daily within a declination of 7° of the Galactic plane. For both surveys, each night a field is observed, it is observed twice, once for each of the g and r bands, and with at least 30 minutes between visits. With these cadences, superoutbursts, whose durations range from a few days to several weeks, are well sampled, as are nova eruptions, high and low states of brightness, and standstills. The g and r bands also provide colour information, a further tool for class separation.

Light curves of cross-matched sources were downloaded from Lasair. Brightness values are given in difference magnitudes, this is the magnitude derived from the positive difference between the flux in the reference image and that in the science image. Where a source contains data points below the reference flux, the difference magnitude light curve profile may deviate from what one would expect for its transient class. Subsequently, these difference fluxes were converted to apparent magnitudes where possible. The formulae used to convert from difference magnitudes to apparent magnitudes and associated errors are given by:

[4] https://sites.astro.caltech.edu/ztf/bts/bts.php

$$m_{\rm corr} = -2.5\ \log_{10}(10^{-0.4\,m_{\rm ref}} + sgn\,10^{-0.4\,m_{\rm diff}}) \tag{6.1}$$

$$\delta m_{\rm corr} = \frac{(10^{-0.8\,m_{\rm diff}}\,\delta m_{\rm diff}^2)^{0.5}}{10^{-0.4\,m_{\rm ref}} + sgn\,10^{-0.4\,m_{\rm diff}}} \tag{6.2}$$

where I simply convert the difference $m_{\rm diff}$ and reference $m_{\rm ref}$ magnitudes to fluxes, sum them considering the sign of the alert (sgn) and convert the results back to magnitude $m_{\rm corr}$. Simple error propagation gives the error $\delta\,m_{\rm corr}$.

To be included in the dataset, two main vetting procedures were followed. The first was to verify the label by checking the references associated with the source. This was easier for the less prevalent classes such as the magnetic systems and AM CVns, where membership can only be verified by means beyond photometry (e.g., spectroscopy, and pulsed X-ray detection), and for dwarf novae further subdivided into the SU UMa and Z Cam classes. For U Gem dwarf novae and those dwarf novae not divided into subclasses, references to literature were less readily available. A second vetting procedure involved inspection of the light curves themselves, where clear misclassifications were identified based on subclass-defining characteristics, and their appropriateness for dataset inclusion could be assessed. In assessing their suitability for inclusion I considered whether phenomena characteristic to a given transient type (e.g., standstills or nova eruption) were present, the number of data points, and whether colour information may be derived. One must be careful to omit examples based on the number of data points, as a limited number may be representative of sources only visible during brightening events. With this consideration in mind, a minimum threshold of at least four points in at least one filter was set.

Example ZTF light curves for each of the classes defined in the following section are given in Figure 6.2. Aside from the usual observing gaps due to the time of year, the limiting magnitude of the telescope in combination with the brightness of the source results in a variety of observational timespans — objects below the limiting magnitude in quiescence may briefly rise into view during episodes of activity, e.g., ZTF22abgglcz

Class/subclass	Number of targets
SU Ursae Majoris	630
Z Camelopardalis	174
U Geminorum	116
nova-like VY Sculptoris	120
nova-like non-VY Sculptoris	123
nova	46
polar	114
intermediate polar	49
AM Canum Venaticorum	46

TABLE 6.1: Number of targets per CV class within the dataset.

and ZTF19aavkbfk. Outbursts of different cycle lengths (time between successive outbursts) are evident for dwarf novae, as are superoutbursts (e.g., ZTF18abosmfh). Evident also are standstills (e.g., ZTF17aaaeepz), long-term changes (high and low brightness states) due to changes in mass-transfer rate (e.g., ZTF18aasncio, ZTF18abcjzao, and ZTF18abryuah), and the various outburst profiles of nova eruptions.

6.2.4 Classification structure

With my task firmly routed in distinguishing between the different types of CV, I settled on a nine-class classification structure that separated the dwarf nova class into their three main subtypes, SU UMa, Z Cam, and U Gem; distinguished between nova-likes and nova-likes containing the VY Scl characteristic; separated the magnetic CVs into their polar and intermediate polar subclasses; with novae and AM CVns making up the remainder. The structure is motivated by the desire for a model that classifies to the highest level of class granularity (to group examples by their most unique traits) while at the same time balancing this desire with the requirement of enough examples to represent the class. This, unfortunately, inhibits our ability to separate the WZ Sge and ER UMa systems from their parent class (SU UMa), and separate novae by their various light curve profiles.

Table 6.1 shows the number of examples per CV class following the vetting procedures. The list is understandably heavily biased towards dwarf novae due to their ubiquity within the CV population.

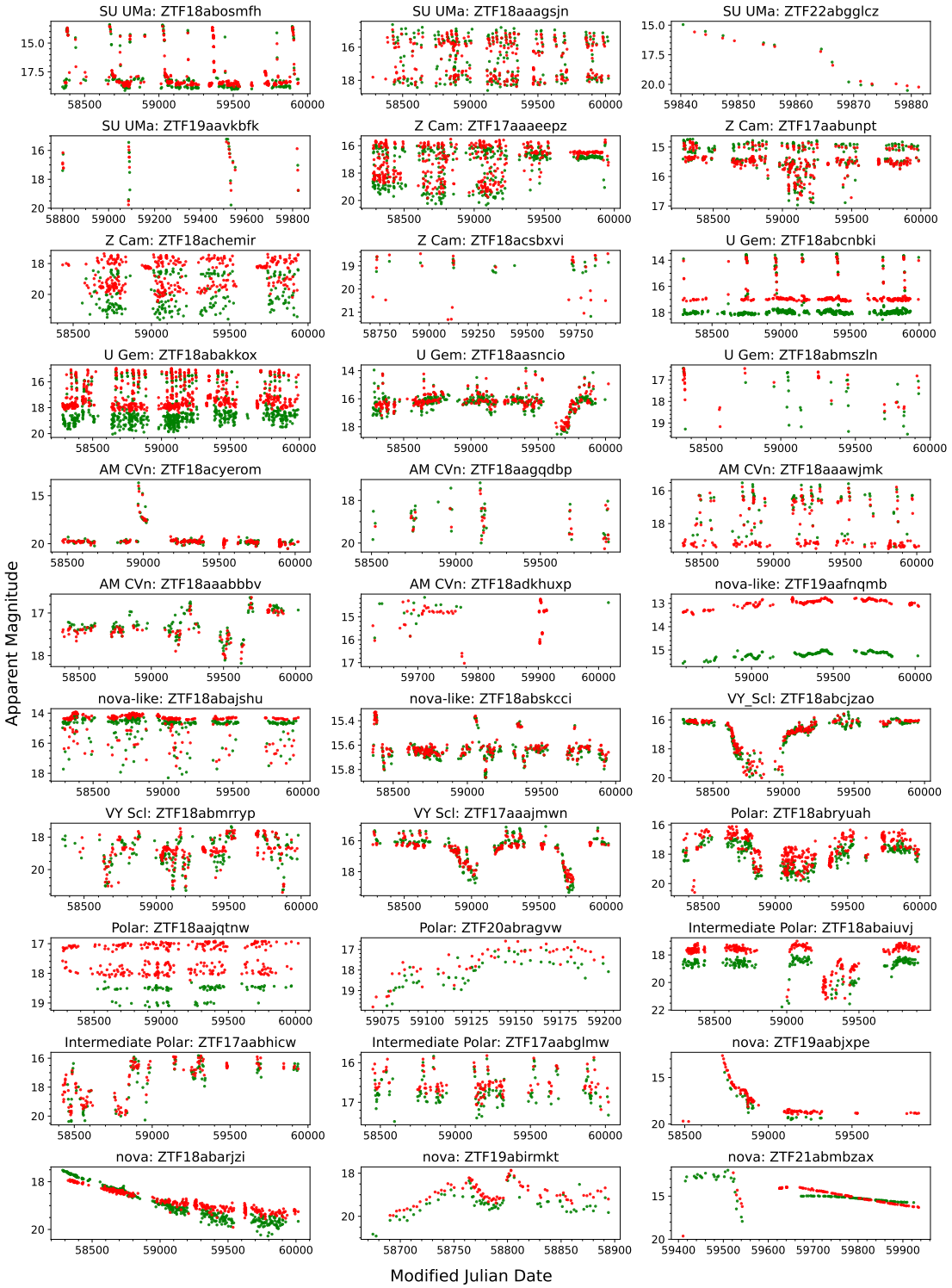


FIGURE 6.2: Example light curves of each CV class. Green and red points indicate g and r band observations respectively.

6.2.5 Features

6.2.5.1 ZTF Light curve derived features

To distinguish between the classes of CV, statistical, percentile and periodicity-based features were extracted from the g and r band source light curves. The suite of features provided by the feATURE eXTRACTOR FOR tIME sERIES (FEETS) python package (Cabral et al., 2018) is comprehensive enough to describe the vast majority of variability characteristics present within the light curves. I therefore make use of them with the addition of several features of my own that are more specifically geared towards CV variability. Non-outbursting systems such as nova-likes and polars are generally well characterised by the FEETS feature set. The diversity of outbursting systems, however, is less well characterised after baseline models revealed the confusion between classes exhibiting such behaviour.

As described in Otulakowska-Hypka et al. (2016), the typical observing cadence, sampling consistency (affected by weather), limiting magnitude and the number of filters that a survey operates under governs the ability to visually recognise and extract features that accurately describe the different types of variability displayed by dwarf nova exhibiting systems. Sub-optimal conditions related to the above inhibit the usefulness of the features extracted. Given the level of classification granularity desired in this work, I developed several simple features that may recognise the presence of phenomena such as superoutbursts, standstills, and their properties.

The *find_peaks* function from the *scipy* Python package locates signal peaks (outbursts peaks in this case) by simply comparing neighbouring brightness values. Not all peaks are identifiable due to undersampled outburst and quiescent phases, and intricacies of the function, though enough useful information is present to obtain the following: an outburst amplitude based on the peak with the largest such value; and rise and decline rates based on the minimum time between outburst peaks and their bases. These features were evaluated for specific outburst amplitude ranges. Recurrence rates are estimated using the frequency corresponding to the maximum power in the Lomb-Scargle periodogram of the light curve. However, it is important to note that the Lomb-Scargle method is not guaranteed to peak at the true recurrence rate if the variability is not strictly periodic. For instance, quasi-periodic signals, irregular recurrence, or stochastic behaviour may

result in a peak at a frequency that does not directly correspond to a recurrence rate. Moreover, the Lomb-Scargle method will output a value even in the absence of strong periodic signals or well-defined outbursts. To distinguish strong periodic signals from weaker or non-periodic behaviour, the ratio of the maximum power to the mean power is used as a diagnostic measure. Additionally, the diverse suite of features extracted aims to capture both periodic and non-periodic variability effectively. With respect to standstills, obvious instances can be characterised by utilising a rolling standard deviation window. Sources with standstills will have windows with high standard deviation values during outbursting periods and low values during standstills. A high ratio of the maximum of the former to the minimum of the latter can detect this dichotomy. This dichotomy, however, is also present in outbursting systems with well-defined quiescent phases (without standstills). One is separated from the other by including the mean brightness level of the window with the minimum standard deviation. A brightness level appreciably higher than the minimum brightness aims to provide the distinction.

Colour is a useful separator of different CV subtypes. In addition to the g-r colour calculated from the average brightness in each filter, I derive the colour for each night where both a g and r band observation was recorded. I include the mean and median of these as features to mitigate the skewing of colour values due to sampling differences between the bands during outburst and quiescence phases. Furthermore, I include the colour at maximum and minimum brightness to account for bluer colours during outbursting phases. All light curve-derived features are given in Tables 6.2 and 6.3.

TABLE 6.2: Features extracted from each of the g and r band light curves. Listed are those available from the FEETS package, where for each a more detailed explanation is provided at `https://feets.readthedocs.io/en/latest/tutorial.html`.

Feature	Description
Amplitude	Half of the difference between the median of the maximum 5% and the median of the minimum 5% magnitudes
AndersonDarling	The Anderson-Darling test is a statistical test of whether a given sample of data is drawn from a given probability distribution (normal distribution)

Autocor_length	Cross-correlation of a signal with itself. Informally, described as the similarity between observations as a function of the time lag between them, useful for finding repeating patterns. Autocorrelation returns a vector, the feature returns the vector length for values less than e^{-1}.
Beyond1Std	Percentage of points beyond one standard deviation from the weighted mean (weighted by the square of the inverse error).
CAR_mean	The mean parameter used to model irregularly sampled time series with the continuous-time autoregressive model (Brockwell & Davis, 2002).
CAR_sigma	The variability parameter used to model irregularly sampled time series with the continuous-time autoregressive model.
CAR_tau	The tau parameter used to model irregularly sampled time series with the continuous-time autoregressive model. Interpreted as the variability amplitude of the light curve.
Con	The number of three consecutive data points that are brighter or fainter than 2σ and normalised the number by N-2.
Eta_e (η^e)	Variability index η is the ratio of the mean of the square of successive differences to the variance of data points.
*FluxPercentileRatioMid**X***	Ratio of centred flux percentile ranges. If $F_{5,95}$ is the difference between the 95th and 5th percentile of ordered magnitudes, then *FluxPercentileRatioMid**X*** $= F_{40,60}/F_{5,95}$, $F_{32.5,67.5}/F_{5,95}$, $F_{25,75}/F_{5,95}$, $F_{17.5,82.5}/F_{5,95}$, and $F_{10,90}/F_{5,95}$, for $\mathbf{X} = 20$, 35, 50, 65, and 80 respectively.
*Freq**i**_harmonics_amplitude_j*	Amplitude of the jth harmonic of the **i**th frequency component of the Lomb Scargle Periodogram

*Freq**i**_harmonics_rel_phase_**i***	The phase corresponding to *Freq**i**_harmonics_amplitude_j* relative to the phase of the first frequency component.
Gskew	Median-of-magnitudes based measure of the skew
LinearTrend	Slope of a linear fit to the light-curve
MaxSlope	Maximum absolute magnitude slope between two consecutive observations
Mean	Mean magnitude
Meanvariance	Ratio of the standard deviation to the mean magnitude
MedianAbsDev	Median absolute deviation of magnitude
MedianBRP	Median Buffer Range Percentage; Fraction (¡= 1) of photometric points within amplitude/10 of the median magnitude.
PairSlopeTrend	Considering the last 30 (time-sorted) measurements of source magnitude, the fraction of increasing first differences minus the fraction of decreasing first differences
PercentAmplitude	Largest percentage difference between either the max or min magnitude and the median
PercentDifferenceFluxPercentile	Ratio of the difference between the 95th and 5th percentile of ordered magnitudes, $F_{5,95}$, over the median magnitude.
PeriodLS	Period corresponding to the frequency of maximum power in the Lomb Scargle Periodogram
Period_fit	The false alarm probability of the largest Lomb Scargle periodogram value.
Psi_CS	*RCS* applied to the phase-folded light curve (generated using the period estimated from the Lomb-Scargle method).
Psi_eta	η^e index calculated from the phase-folded light curve
Q31	Difference between the third and first quartile of the light curve magnitudes

Rcs	Range of a cumulative sum (R_{CS}) of the light curve. Defined as: R_{CS} = max(S) — min(S), where $S = \frac{1}{N\sigma}\sum_{i=1}^{l}(m_i - \bar{m})$. N represents the number of points, with $i = 1, 2, ..., N$.
Skew	Skewness of the magnitudes
SlottedA_length	Slotted autocorrelation length — same as *Autocor_length* except that time lags are defined as intervals or slots instead of single values
SmallKurtosis	Small sample kurtosis of the magnitudes.
Std_g	Standard deviation of magnitudes.
StetsonK	Robust measure of the kurtosis (Stetson, 1996).
StetsonK_AC	Variability index derived based on the autocorrelation function of each lightcurve (Stetson, 1996).
Q31_colour	*Q31* applied to the difference in the g and r band magnitudes.
StetsonJ	A robust version of the Welch/Stetson variability index I (Stetson, 1996) describing the synchronous variability of different bands.
StetsonL	Variability index describing the synchronous variability of different bands that utilises both *StetsonJ* and *StetsonK*.

TABLE 6.3: Additional light curve derived features implemented in this work.

Feature	Description
median	Median of magnitudes.
min_mag	Minimum magnitude (maximum brightness).
max_mag	Maximum magnitude (minimum brightness).
n_obs	Number of light curve data points.
dif_min_mean	Difference between minimum and mean magnitude.
dif_min_median	Difference between minimum and medium magnitude.
dif_max_mean	Difference between maximum and mean magnitude.
dif_max_median	Difference between maximum and median magnitude.

dif_max_min	Absolute difference between maximum and minimum magnitude.
temporal_baseline	Duration of the light curve.
pwr_max	Maximum power of Lomb Scargle periodogram.
pwr_maxovermean	Maximum over the mean power of the Lomb Scargle periodogram of the light curve.
*npeaks_**X**to**Y***	Number of peaks with amplitude between **X** and **Y**. $\mathbf{X} \in (0.5, 1, 2)$ and $\mathbf{Y} \in (1, 2, 5)$. *npeaks_above5* for peaks above 5 magnitudes.
*rrate_**X**to**Y***	Maximum rise rate of peaks with amplitude between **X** and **Y**.
*drate_**X**to**Y***	Maximum decline rate of peaks with amplitude between **X** and **Y**.
*amp_**X**to**Y***	Maximum amplitude of peaks with amplitude between **X** and **Y**.
*rollstd_ratio_t**A**s**B***	Calculate the rolling standard deviation of the light curve with a window size $\mathbf{B} \in (5, 10)$, where the threshold for the minimum light curve data points, $\mathbf{A} \in (10, 20)$, is met. The ratio of the highest to lowest standard deviation of these windows is the output.
*stdstilllev_t**A**s**B***	Ratio of the mean magnitude of the window with the lowest standard deviation to the magnitude range of the light curve — i.e., standstill location relative to the maximum brightness.
*pnts_leq_rollMedWin20-**C**mag*	Number of data points within a rolling window of 20 observations that are brighter than **C** magnitudes of the median magnitude of that window, where $\mathbf{C} \in (1, 2, 5,)$.
*pnts_geq_rollMedWin20-**D**mag*	Number of data points within a rolling window of 20 observations that are fainter than **C** magnitudes of the median magnitude of that window, where $\mathbf{D} \in (1, 2, 3)$.
*pnts_leq_median-**E**mag*	Number of data points brighter than **E** magnitudes of the median magnitude of the light curve, where $\mathbf{E} \in (1, 2, 5)$.

*pnts_geq_median-**F**mag*	Number of data points fainter than **F** magnitudes of the median magnitude of the light curve, where $\mathbf{F} \in (1, 2, 3)$.
clr_mean	Mean of the colours derived at each epoch (night) where an observation in both the g and r band was obtained. Where no epochal colour information is available for a source, the difference between the mean g magnitude and mean r magnitude is used.
clr_median	Same process as used to calculate *clr_mean*, this time with the median instead of mean magnitude.
clr_std	Standard deviation of the epochal colour.
clr_bright	Colour obtained from the epoch where the system is at its brightest. Where epochal colour is unavailable, this is the difference between the minimum g and r band magnitudes.
clr_faint	Colour obtained from the epoch where the system is at its faintest. Where epochal colour is unavailable, this is the difference between the maximum g and r band maximum.

6.2.5.2 Features derived from Gaia

In addition to the light curve-based features, data from Gaia DR3 (Gaia-Collaboration et al., 2022) is incorporated. Specifically, I utilised photometry from the G band, red photometer (RP), and blue photometer (BP) filters, including colour indices and astrometric data such as parallax and proper motion, along with their associated uncertainties. Distances and absolute magnitudes were also derived; however, their errors were excluded as they were deemed redundant, given the inclusion of flux and parallax uncertainties. These supplementary data are included as features that are described in Table 6.4. Such metadata are not available for every source, and one would not expect this information to be available for new sources of unknown class that one wishes to classify. I discuss this issue in subsection 6.2.9.

TABLE 6.4: Supplementary data from Gaia EDR3 incorporated as dataset features

Feature	Description
ra, dec, ra_error, dec_error	Right ascension, declination, and associated standard errors
l, b	Galactic longitude and Galactic latitude
ecl_lon, ecl_lat	Ecliptic longitude and Ecliptic latitude
bp_rp, bp_g, g_rp	BP-RP, BP-G, and G-RP colours
*phot_**X**_mean_flux*	Mean flux in the G, integrated BP, or integrated RP bands — corresponding to $\boldsymbol{X}$ = *g, bp, or rp* respectively
*phot_**X**_mean_flux_error*	Error on the mean flux in the **X** band
*phot_**X**_mean_mag*	Mean magnitude in the G, integrated BP, or integrated RP bands — corresponding to $\boldsymbol{X}$ = *g, bp, or rp* respectively
parallax, parallax_error	Gaia parallax in milliarcseconds (mas) and standard error
pm	Proper motion (mas/year)
pmra_error, pmdec_error	Standard error of the proper motion in right ascension and declination directions (mas/year)
phot_g_n_obs, phot_bp_n_obs, phot_rp_n_obs	Number of observations in the Gaia G, BP, and RP bands.
phot_g_mean_mag, phot_bp_mean_mag, phot_rp_mean_mag	Mean magnitude in the Gaia G, integrated BP and RP bands
distance	Distance to the source derived from the inverse parallax (parsecs)
absmag_g, absmag_BP, absmag_RP	Absolute Gaia G, integrated BP and RP magnitudes derived from parallax.
nu_eff_used_in_astrometry	Effective wavenumber of the source. Calculated as the photon-weighted inverse wavelength, calculated from the BP and RP spectra ($\lambda^{-}1$).

6.2.6 Training, validation and test sets

The size of the dataset is insufficient for a separate validation set, with minority class examples numbering only a few dozen. I therefore opt for a technique designed for such cases, stratified k-fold cross-validation. This technique allows me to maintain an adequately sized training set and serves to assess the consistency of the model (and data). I use a stratified train-test set split ratio of 70:30 and a 10-fold stratified cross-validation (on the training set) procedure for hyperparameter tuning and model evaluation. The 70:30 split holds back for testing at least a dozen examples for minority classes whilst providing a high proportion of examples for the algorithm to learn patterns during training and for validation.

6.2.7 Feature selection

The dataset consists of over 250 features, and with only 1,439 examples, one introduces the 'curse of dimensionality' (Bellman, 1957), which refers to a set of problems arising from high-dimensionality datasets. As you add dimensions (features) you rapidly increase the minimum amount of samples required to adequately represent all combinations of feature values in your dataset. Increasing the dimensionality increases the complexity of the model whilst also causing the model to become increasingly dependent on the training set, thus leading to overfitting. Selecting the features most informative for the task enables ML algorithms to train faster, reduces complexity allowing for easier interpretation, reduces overfitting, and can improve model accuracy for the right subset of features. To identify the optimal feature subset, the Variance Inflation Factor (VIF; Vu et al. 2015), the one-way Analysis Of Variance (ANOVA; Quirk 2012), and the mutual information score (MI; Quirk 2012) methods were examined from the filter feature selection family that measures the relevance of features by their correlation with the dependent variable. From the wrapper method family, which examines the usefulness of a subset of features by training a given model on them, the forward feature selection (FFS) method was chosen. These methods were applied to the training set only to avoid data leakage — information about the target being present in the training set that would not be available when the model is used for prediction (Singhi & Liu, 2006; Demircioğlu, 2021).

FFS is utilised for all but the Decision Tree-based algorithms as they naturally determine the most important features during the tree-building process. VIF is particularly beneficial when dealing with feature redundancy that may arise when a feature is derived from both the g and r bands. We experimented with VIF values of 10, 5, 2.5, and 1.5 for all but the Decision Tree-based algorithms since decision trees select features in a greedy fashion and make no assumptions about relationships between features.

One-way ANOVA was used to identify the significance of each feature ordered by p-value. A given algorithm was then trained using the top x% of the most significant features and the model cross-validation performance was recorded. This step was repeated, increasing the values of x in 5% increments from 5% to 95%, to arrive at a subset of features where model performance was strongest. This method is akin to forward feature selection, though with features added based on a statistical test rather than

overall model performance. The motivation for the usage of one-way ANOVA lies in its goal to select a set of features that hold significant importance in differentiating between classes. Similar to FFS and VIF, this method was applied exclusively to algorithms that do not utilise Decision Trees.

Under the MI feature selection protocol, the most performant features were identified in the same way as for one-way ANOVA, resulting in slight variations in the optimal subset of features for each algorithm. In a similar fashion to one-way ANOVA, MI aims to select features most crucial for class distinction. However, MI quantifies the information shared between features and the outcome, thereby unveiling non-linear, intricate relationships. This feature reduction method was not employed for the Decision Tree-based algorithms for the same reasons as above, and furthermore, MI is at the heart of the operation of these algorithms.

6.2.8 Class balancing

Large disparities in the number of examples for each class significantly affect the performance of a model. Differences in class prevalence cause algorithms to be biased towards learning patterns more specific to the majority class and produce models that perform poorly in minority class predictions. To handle the class imbalance present within the dataset (see Table 6.1) I tested both a non-sampling method, class weighting (should the algorithm permit such a strategy), and random undersampling of the majority class combined with the minority class over-sampling technique Adaptive Synthetic (ADASYN; Haibo et al. 2008), a variation of Synthetic Minority Over-sampling Technique (SMOTE; Chawla et al. 2002).

6.2.9 Missing Data

Missing data due to insufficient data points during the light curve feature extraction process accounts for as much as 20% for a given feature. Whilst that due to unavailability of metadata accounts for up to 33%. Many machine learning algorithms do not support missing values, therefore strategies must be implemented to address this absence of data. The most common and simplest strategy is to replace (or impute) missing values with the mean or median of the feature, however, this method ignores relationships between

features and reduces the variance of the variable, thereby introducing bias to the model. The following approach aims to mitigate such bias.

Firstly, the reasons for missingness are assessed. Depending on the context, I assign either the value from the other filter (if available), a contextually relevant substitute (e.g., the overall mean $g - r$ colour, defined as the mean g-band magnitude minus the mean r-band magnitude, if the epochal mean colour (*clr_mean*) is missing), or the feature value is left as missing. For these remaining missing values I utilise the Scikit-learn implementation of the K Nearest Neighbour imputation method (Troyanskaya et al., 2001). For each sample, each missing feature is imputed using the average of the values (weighted-by-Euclidean distance) from the K nearest neighbours in feature space where that feature value is present.

6.2.10 Machine Learning algorithms

The algorithms whose performance I evaluate are Scikit-learn's (Pedregosa et al., 2011) Python implementations of Random Forest (RF; Breiman 2001), K-Nearest neighbours (KNN; Zhang 2016), Gaussian Naive Bayes (GNB; Zhang 2004), and Linear Discriminant Analysis (LDA; Hastie et al. 2003). Also used are the Extreme Gradient Boosting (XGBoost) algorithm (Chen & Guestrin, 2016) and Keras (Chollet, 2021) implementation of an Artificial Neural Network (NN) in the form of a Multi-Layer Perceptron — a fully connected multi-layer NN (Kruse et al., 2022). Furthermore, for model evaluation and interpretability purposes, I used the Gaspar (2018) Python implementation of Generative Topographic Mapping (Bishop et al., 1998).

The array of algorithms embodies a diverse spectrum of classification strategies chosen to extract optimal insights from the dataset. RF is adept at navigating intricate patterns in data through its ability to handle non-linear relationships, high-dimensional data, and noisy features. XGBoost is known for delivering high-performance scalability, often surpassing other algorithms and underscoring the potential of ensemble methods. KNN adds instance-based learning to the mix, GNB adds probabilistic modelling, and LDA is adept at discerning linear separability. Meanwhile, the multi-layer perceptron is a fundamental deep learning architecture, these are capable of capturing intricate patterns in data.

6.3 Results

6.3.1 Classifiers

In this study, to distinguish between the nine CV classes, I evaluated several algorithms: Gaussian Naive Bayes, Linear Discriminant Analysis, K-Nearest Neighbors, Random Forest, XGBoost, and a multi-layer perceptron neural network. To address the class imbalance, I used either the class weighting method (where possible) or the ADASYN oversampling technique in combination with random undersampling to balance the training set. The training was conducted on subsets of features determined through the mutual information score, variance inflation factor, the one-way ANOVA method, or forward feature selection. I assessed the resultant models based on overall accuracy, macro averages of precision, recall (equivalent to balanced accuracy for the macro average), and F1-score. These are provided in the heatmap shown in Figure 6.3 within the first four columns. The corresponding precision, recall, and F1-scores for each class are provided in the remaining columns.

To compare the test set performance metric means of different classifier groups, I conducted T-tests. The results indicate that GNB and KNN-based classifiers performed poorly on the test set compared to the other algorithms (F1-score of 0.44 ± 0.04 and 0.54 ± 0.04, respectively, with a p-value of $p = 8.6 \times 10^{-11}$ at $\alpha = 0.05$). However, there was no significant performance difference when using over/under sampling compared to class weighting (or no such method) ($p = 0.07 - 0.86$ for all metrics). Regarding feature reduction methods, I observed small but not significant performance improvements when using the one-way ANOVA and mutual information, while the use of variance inflation factor led to a performance drop.

The class-specific performance associated with each model revealed difficulties in correctly classifying the AM CVn and intermediate polar classes, irrespective of the algorithm used or any effort to address the class imbalance. These two classes, along with the nova class, have the lowest sample size. Despite the small sample size, the light curves (and metadata) of the novae are sufficiently distinct for the algorithms (especially NN, RF, and XGBoost) to distinguish them from the remaining classes. All models, except for GNB, performed well in classifying the SU UMa class.

TABLE 6.5: Top 5 ranked classifiers based on the macro-averaged F1-score. Listed are the algorithm, the method used to handle class imbalance and the method used to reduce the number of features. The class balancing methods are abbreviated as SMPL, WTD, or -, depending on whether over/under sampling methods, class weighting, or no class balancing method was implemented, respectively. The only feature selection methods in this list are those abbreviated as MUI or -, for mutual information or no feature selection method (full list of features used), respectively.

Rank	Algorithm	Imbalance	Feature selection	F1-score
1	XGB	—	—	0.62
2	RF	SMPL	—	0.58
3	XGB	SMPL	—	0.57
4	NN	WTD	MUI	0.57
5	LDA	—	—	0.57

To select the model for the pipeline, I based the decision on the macro F1-score with some consideration for the performance on the lowest sample size classes. Table 6.5 presents the top 5 models based on the macro F1-score, while Figure 6.4 shows the class-specific 'p-value table' resulting from a McNemar's test for each pair of these models. The figure indicates no significant prediction disagreements between these algorithms for the SU UMa, nova, and intermediate polar classes. However, models ranked in the top 3 show significant prediction disagreements compared to models ranked 4 and 5 with regards to Z Cam. For the AM CVn class, the XGBoost classifier, implemented without explicit class balancing or feature reduction, significantly outperformed the other models. As a result, I selected this XGBoost model trained with 500 decision trees at a maximum tree depth of 14, as the classifier for the second stage of the pipeline.

It should be noted that certain aspects of the model selection, such as the variation of examples apportioned to the training, validation, and test sets, the NN weights initialisation, and the feature selection for each tree of the RF and XGBoost models, were randomly selected. Thus, different random initialisations could have led to the selection of any of the models generated from the NN, RF, and XGBoost algorithms.

6.3.2 Performance

The per-class performance of the model as implemented on the test set is described in Table 6.6, while the corresponding confusion matrix is shown in Figure 6.5. Evident are the following. SU UMa is responsible for the highest precision and recall scores, contributing greatly towards an increase in the overall classification performance, Z Cam and VY Scl are also well picked out by the classifier. The overall performance suffers

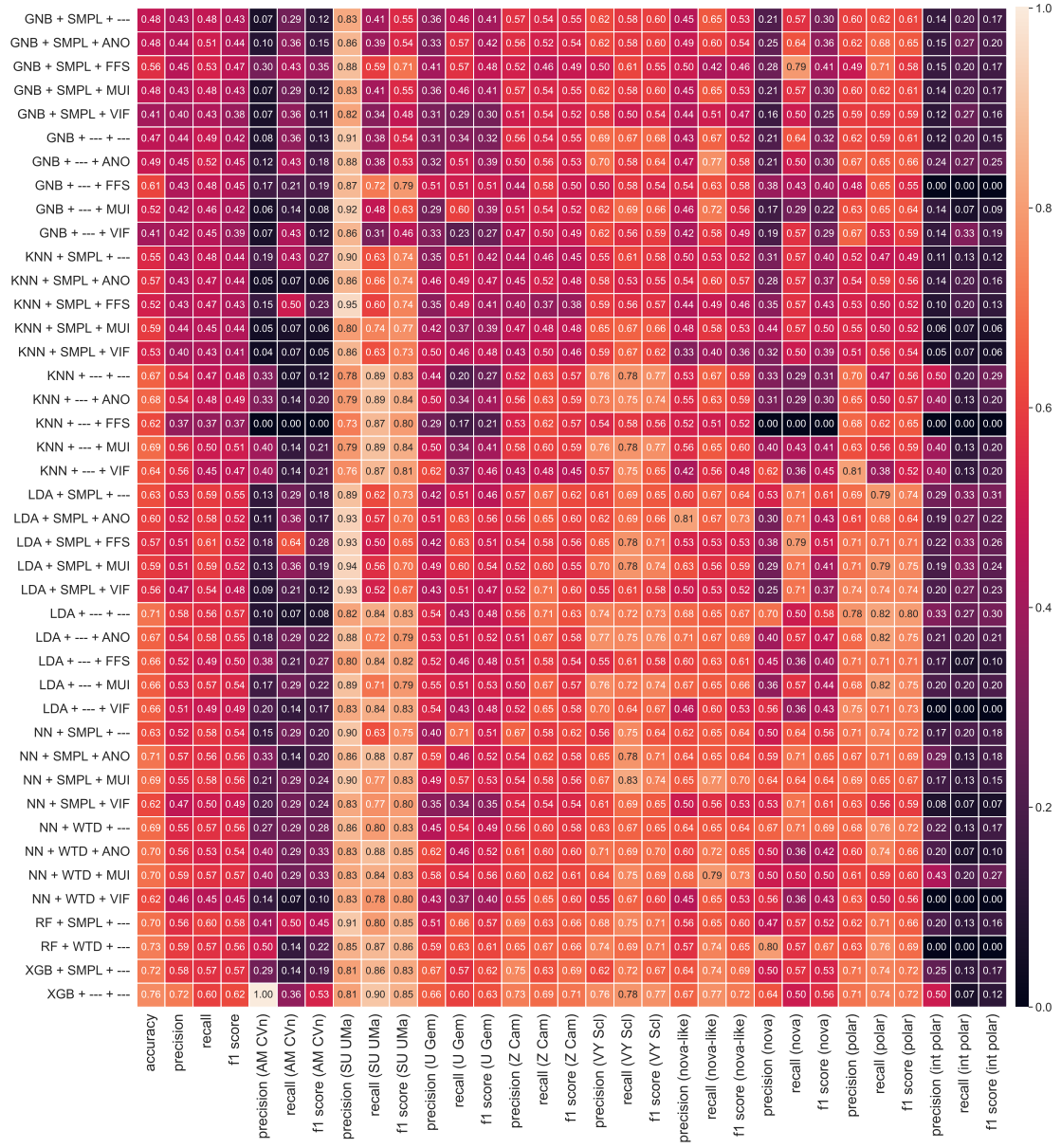


FIGURE 6.3: Presented as a heatmap are, the accuracy, and the macro average quantities of precision, recall, and F1-score for each classifier variant. Alongside these are the precision, recall, and F1-score for each class. Classifiers are labelled as follows: classifier + class balancing method + feature selection method. Classifier abbreviations are as described in the text, the class balancing methods are abbreviated as SMPL, WTD, or —, depending on whether over/under sampling methods, class weighting, or no class balancing method was implemented, respectively. Feature selection methods are abbreviated as ANO, FFS, MUI, VIF, or —, for one-way ANOVA, forward feature selection, mutual information, variance inflation factor, or no such implementation (full set of features used), respectively.

FIGURE 6.4: The per-class p-values from McNemar's tests were conducted between each pair of the top 5 ranked classifiers from Table 6.5. For ease of reference, these are, from rank 1 to 5, XGB + — + —, RF + SMPL + —, XGB + SMPL + —, NN + WTD + MUI, AND LDA + — + —. The significance threshold is set to p=0.05, and the classifier descriptions and abbreviations are as described in the caption of Figure 6.3.

noticeably due to the performance of the intermediate polar class. Intermediate polars represent a class subject to one of the largest amounts of training set oversampling, due to a low number of examples.

Also falling within this high oversampling bracket are the AM CVns and novae. Despite this, they are responsible for strong precision scores such that 100% of examples predicted as AM CVn and 64% of examples predicted as nova are true members of the class. However, this does come at the expense of lower recall scores, 0.36 for AM CVns and 0.50 for novae. Those true AM CVn members that are misclassified are mostly assigned the SU UMa class, as are true members of the nova class.

TABLE 6.6: Classification report for the XGBoost model. For each class of CV the precision, recall, F1 score, and the number of test set examples are given. The macro average (or arithmetic mean) of each metric, accuracy and balanced accuracy are also provided.

Class	Precision	Recall	F1 score	Test set amount
AM CVn	1.00	0.36	0.53	14
SU UMa	0.81	0.90	0.85	189
U Gem	0.66	0.60	0.63	35
Z Cam	0.73	0.69	0.71	52
Intermediate Polar	0.50	0.07	0.12	15
Nova	0.64	0.50	0.56	14
Nova-like	0.67	0.77	0.72	43
Nova-like VY Scl	0.76	0.78	0.77	36
Polar	0.71	0.74	0.72	34
Macro average	0.72	0.60	0.62	432
Accuracy			0.76	432
Balanced accuracy			0.60	432

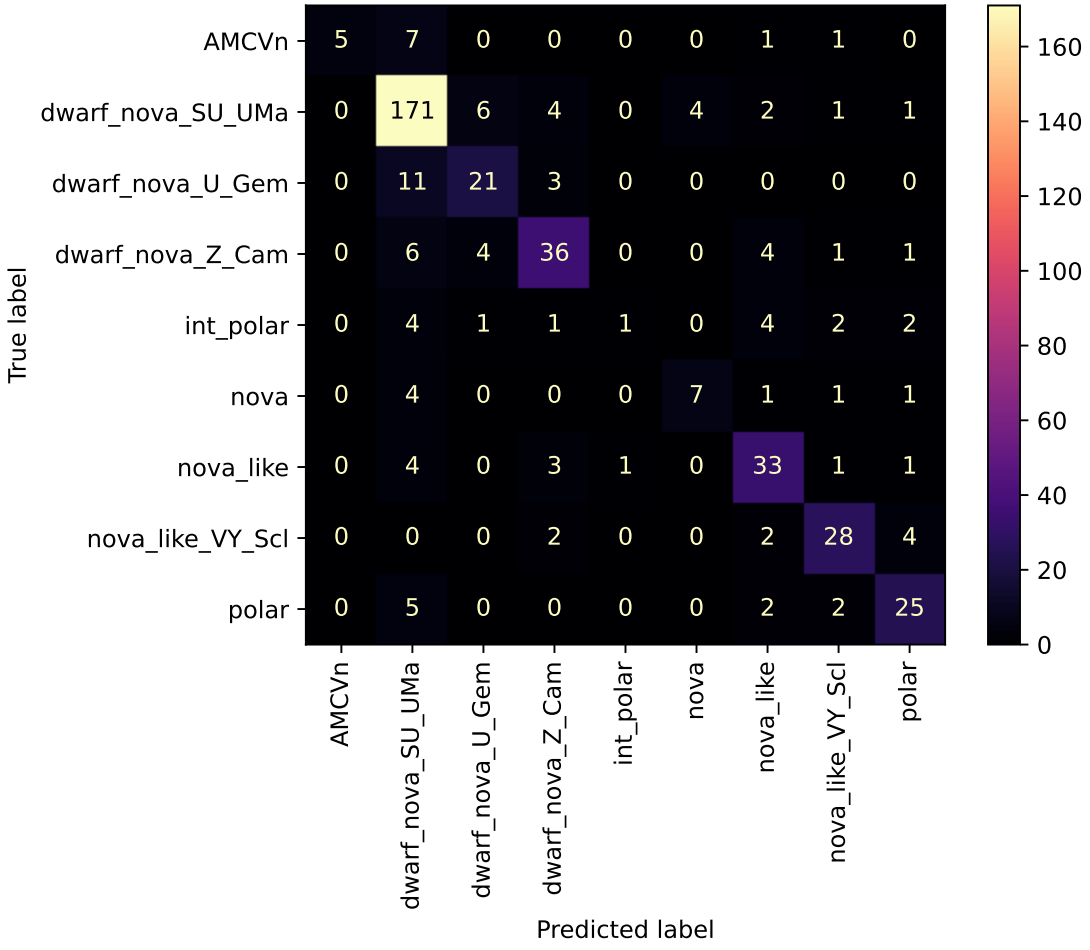


FIGURE 6.5: Confusion matrix for the XGBoost model.

The classifier performs well in distinguishing between systems that regularly display dwarf nova outbursts (where we exclude intermediate polars) from those that do not. Should we group those classes into those that exhibit these outbursts and those that do not, the precision and recall scores for the dwarf nova exhibiting class would be 0.92 and 0.94 respectively, while for non-dwarf nova exhibiting systems, 0.88 and 0.83. Confusion between dwarf nova exhibiting systems is an area where the model performance suffers. Notable is the mislabelling of AM CVn members as SU UMa; and the contamination of predictions of the U Gem class by SU UMa and Z Cam members. Similarly, confusion between non-dwarf nova exhibiting systems also plays a factor: true intermediate polar members are confused for nova-likes, VY Scl and polars; and confusion between the nova-like, VY Scl and polar classes is present. Reverting the description of performance back to the 9 class problem, notable is the significant misclassification of true Z Cam members with the nova-like class and the significant contribution of false positives by the SU UMa class towards the predictions of the nova class.

With respect to the ROC Curves (Figure 6.6), in all cases, the classifier performed much better than a random guess, depicted by the 'chance level' line. An AUC score above 0.93 for all but the intermediate polar and AM CVn classes represents a strongly performing classifier, where the resultant micro and macro averages are 0.96 and 0.92. This is a further illustration of the findings within the confusion matrix and classification report.

The importance of each feature for DT-based models can be given by the feature importance scores. The 20 features with the largest effect on the model's predictive accuracy are plotted in Figure 6.7. Ranked highest is the Gaia RP band absolute magnitude (*abs_mag_rp*); Gaia BP and G absolute magnitudes also feature within the list. ZTF and Gaia colours feature strongly, with the brightest epochal colour (*clr_bright*), Gaia G-RP and Gaia BP-RP colours within the top 10. The slope of a linear fit to the ZTF r band light curve is deemed highly relevant for classifier performance, as is the auto-correlation length in the ZTF g band. Periodicity-based features within the list come in the form of the frequency of maximum power in the Lomb Scargle periodogram of the r band light curve. Features for identifying outbursts are represented by the number of points brighter than the rolling median. Features that test for the synchronous light curve variability across both bands come in the form of *StetsonJ* and *StetsonL* (see Table 6.2). The list therefore contains a mixture of features that cover periodicity, photometry, and statistical descriptors.

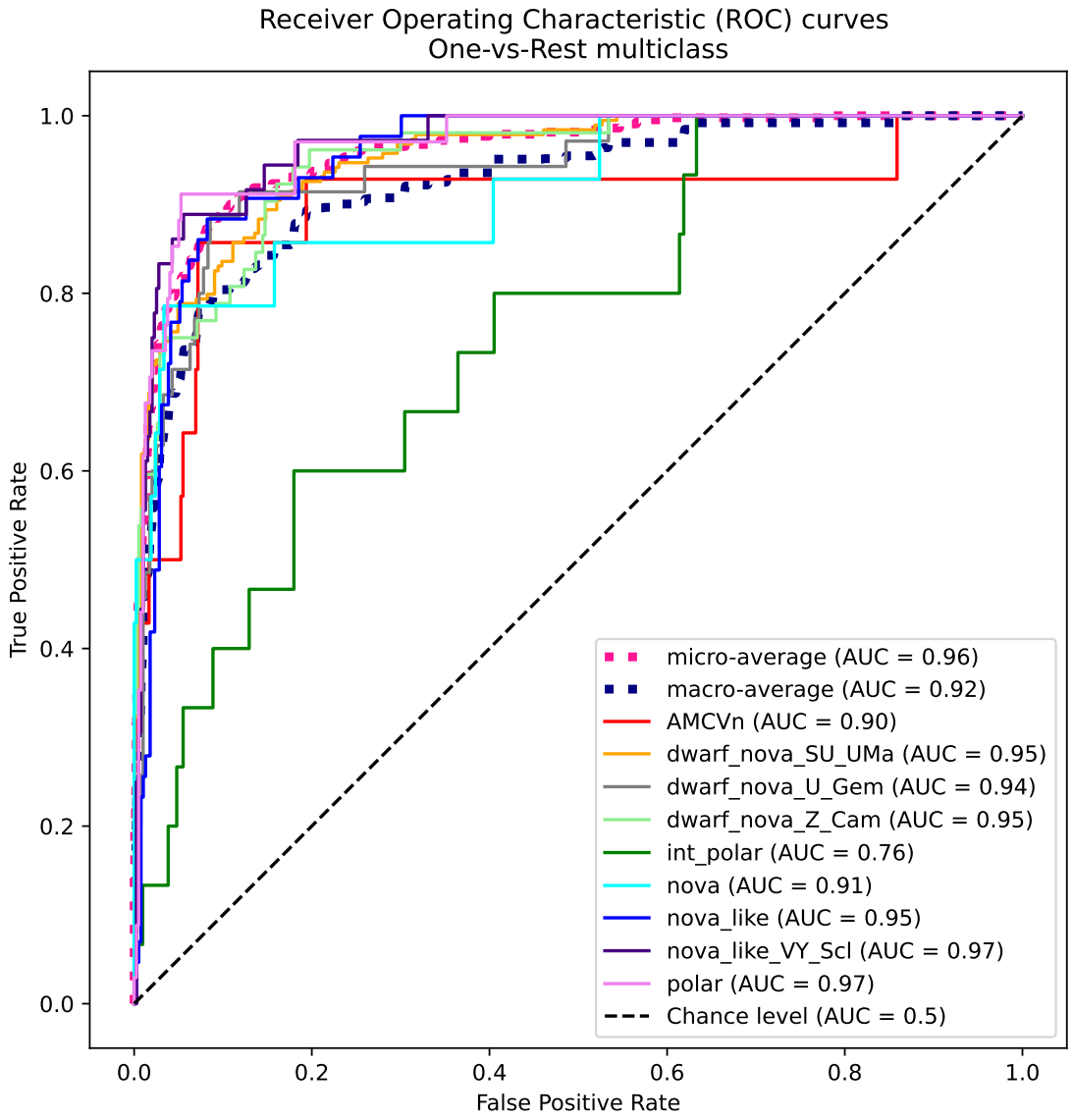


FIGURE 6.6: Receiver operating characteristics for the XGBoost classifier.

6.3.3 GTM Latent space representations

6.3.3.1 GTM for model assessment and feature relevance

One may utilise Generative Topographic Mapping to evaluate the ability of a classifier to distinguish between classes and to identify the features responsible for the assignment of a given class rather than an overall feature importance list that only provides the features responsible for overall model performance. To do this I input the posterior class probabilities for the training set output by the classifier into the GTM framework. Therefore the data space is a class probability space of nine dimensions. Each example from the training set will have posterior probabilities of belonging to each class evaluated by the

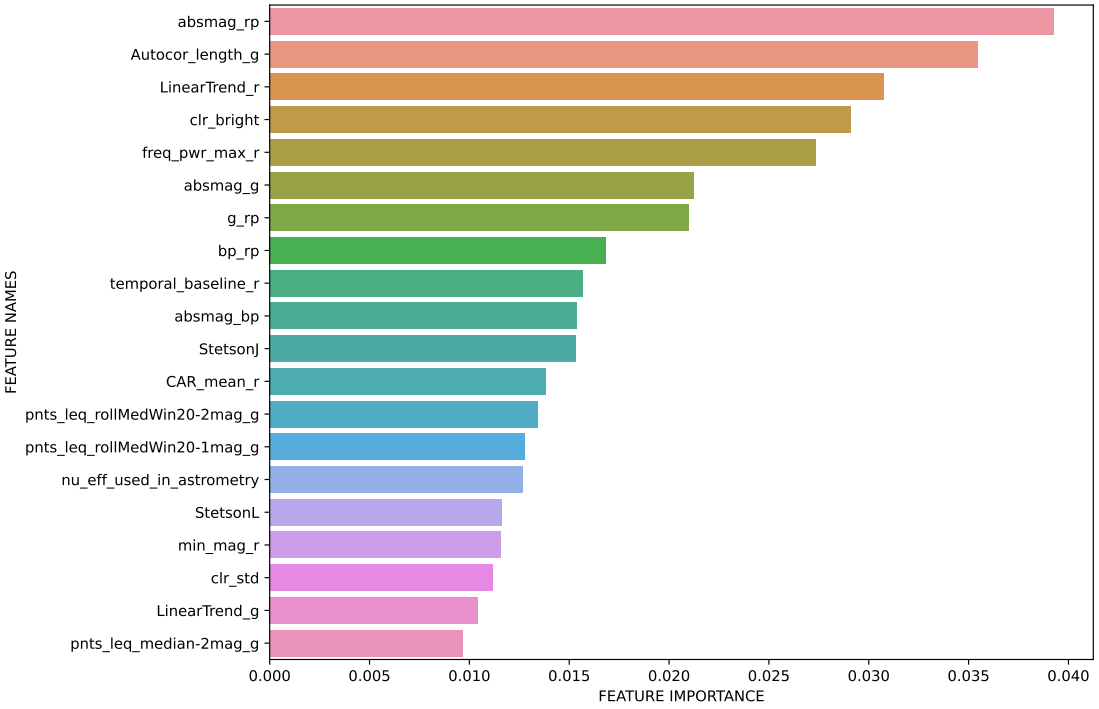


FIGURE 6.7: Feature importance scores for the 20 most influential features within the chosen classifier model. Feature importance refers to a class of techniques for assigning scores to input features to a predictive model, indicating the relative importance of each feature when making a prediction.

classifier, these probabilities define their location in class probability space. Distinct clusters of these examples located in regions with high probability along a particular probability space dimension would represent a classifier that can accurately distinguish between classes. Since these clusters define the Gaussian centres, they are mapped to the corresponding nodes in latent space. One may then evaluate this class separability within the latent space representation. This representation forms a grid of squares, each defining a node, colour-coded based on the location of the associated probability space Gaussian centre along a given probability space axis (or particular class probability) — these are referred to as class maps.

For feature responsibilities, I simply average a particular feature value for all examples assigned to a given node, 'assigned' meaning the node with the highest likelihood of being responsible for a given example. The average for each node can then be used to produce a 2D histogram consisting of the same above latent space grid with squares colour-coded by these averages, one for each feature. The distribution of mean feature values can be analysed against the distribution of classes in the class maps to identify class-specific features.

6.3.3.2 Class and feature maps

Class maps generated using GTM, are presented in Figure 6.8. These latent space representations of class probability space structures assess the class separability of our ML model. The class maps clearly show the existence of structures that are located in fairly distinct regions, each associated with a particular class. This is representative of a classifier that has effectively learnt patterns within the data necessary for class distinction. These structures are extended, with their cores represented by the highest probability of belonging to the associated class, whilst as we move away from the cores, the probabilities diminish (represented by the colour scale). Structures extend into regions associated with that of other classes, indicating some class confusion, thus reflecting observations within the confusion matrix. The highest class probabilities are associated with the SU UMa, U Gem, Z Cam, nova-like, and VY Scl classes — their structure cores exceed 0.80 in class prediction probability. Structures for the AM CVn, nova and polar classes are also present, though with class probabilities no higher than 0.7 and 0.8 respectively. As mirrored in the confusion matrix, the intermediate polar structure, though located in a relatively distinct region, is only responsible for a core class probability of 0.62.

Another interesting feature of the maps is that outbursting systems tend to reside along the top edge and down the left edge, while systems that are not expected to display dwarf nova outbursts are located along the right and bottom edges of the maps. This concurs with the observation of the effectiveness of the model in distinguishing outbursting from non-outbursting systems. The nova class is the only one located away from any edge.

The most obvious blending between structures (or equivalently, confusion between classes) is evident for the SU UMa class — the most prevalent class in the dataset. Its structure extends well into the AM CVn and U Gem regions, also coming into contact with Z Cam and nova. Z Cam is responsible for a well-defined structure (top right) that extends into nova-like class probability space, and a tenuous one ($\sim$0.2 in Z Cam class probability) that is more strongly associated with the nova-like, VY Scl and intermediate polar classes. Nova-likes are also responsible for a tenuous, secondary structure (bottom right) more strongly associated with intermediate polars. There is also a clear overlap between nova-like, VY Scl and intermediate polar classes, and structure blending is evident between AM CVn, nova, and polar classes.

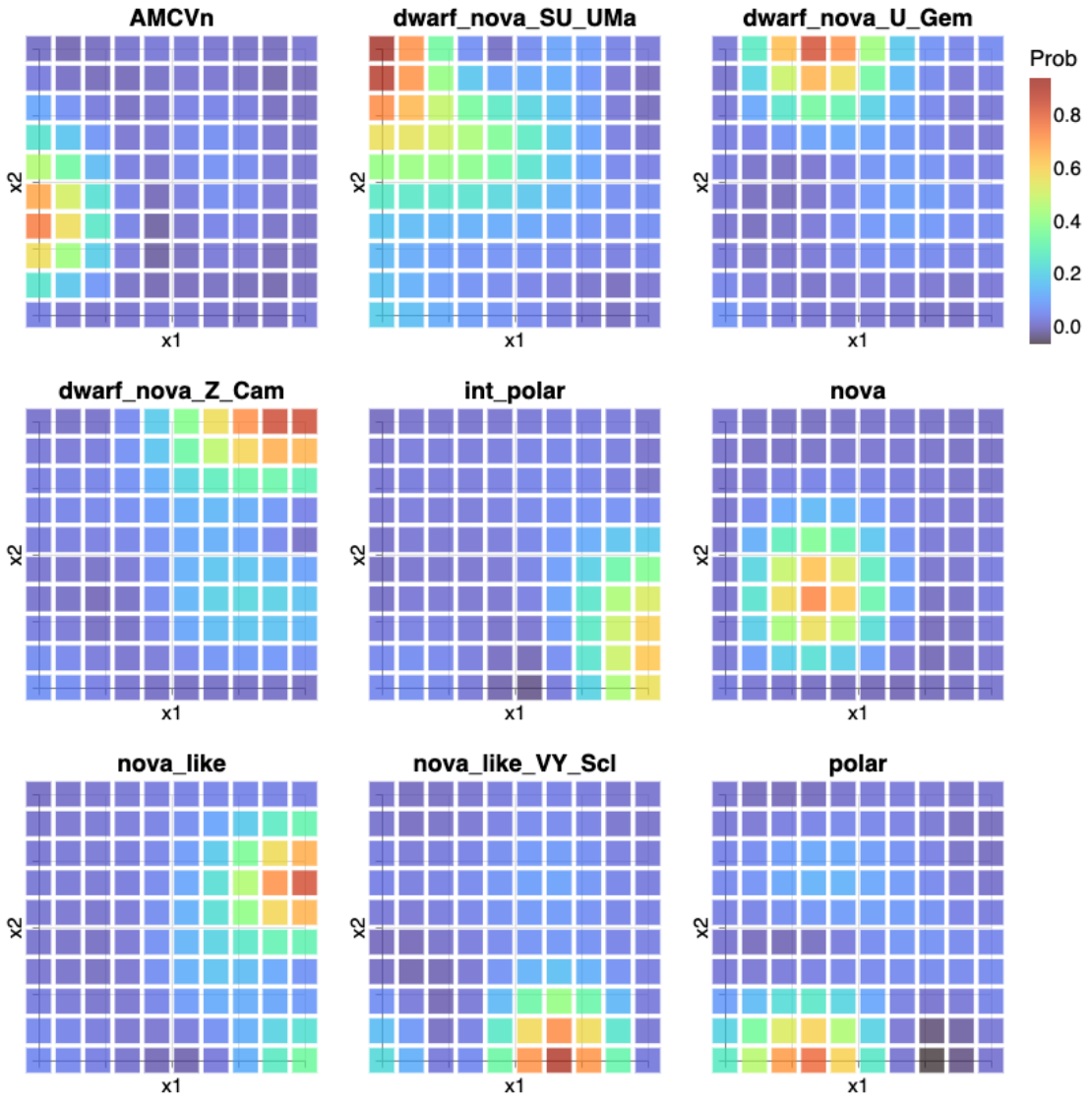
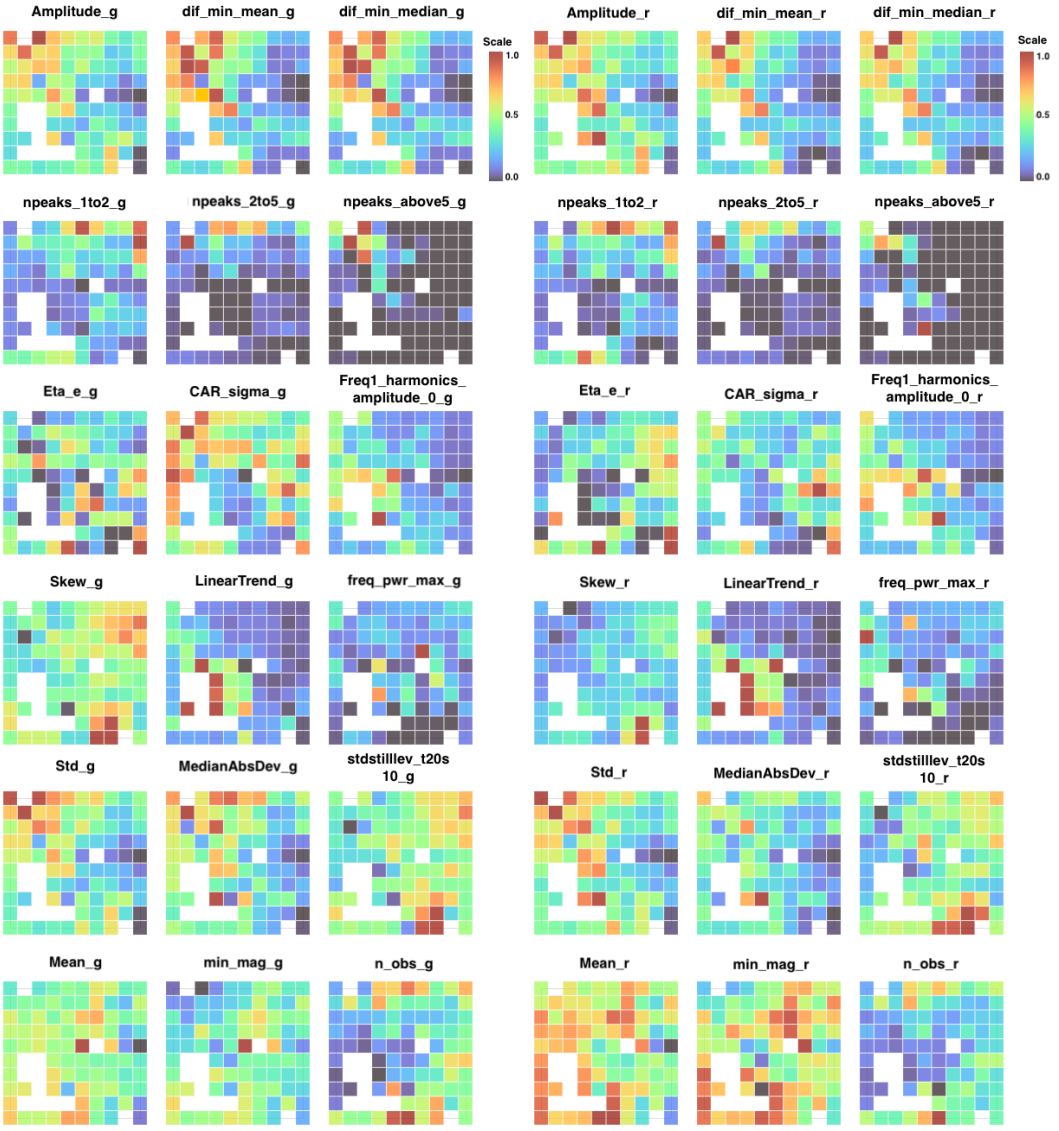


FIGURE 6.8: GTM latent space visualisation of the class posterior probability space from the XGBoost classifier chosen for the pipeline.

Figures 6.9a and 6.9b are a selection of feature maps for features derived from the g and r band light curves. Several further feature maps are shown in Figure 6.10 representing features derived from a combination of the g and r band light curves and Gaia DR3. They represent the average feature values of examples assigned to each of the latent space nodes. The feature maps can be used as tools to identify the features most responsible for the assignment of a given class. This is done by comparing class map structures with those within the feature maps. While examination of the feature maps is reserved for the discussion section, it is clear that structures and patterns exist within them that coincide with class-map structures. For example, high values for amplitude and variability-based features (e.g., *Amplitude*, *Std*, *MedianAbsDev*, and *npeaks*) correspond



(A) ZTF-g band feature maps (B) ZTF-r band feature maps

FIGURE 6.9: GTM-generated feature maps for the XGBoost model. Compare high and low-value regions to class maps to pinpoint key features for class assignment. White squares indicate empty nodes, to which no examples are assigned, determined by node responsibility.

to outbursting systems; the fewest number of data points, *n_obs*, are associated with AM CVn, SU UMa and nova classes; and the bluest colours are associated with the AM CVn class.

6.3.4 Alert stream pipeline

With the aim of the alerts filter to minimise the number of possible non-CVs and maximise potential CVs, this was best achieved with the following procedure. The Sherlock

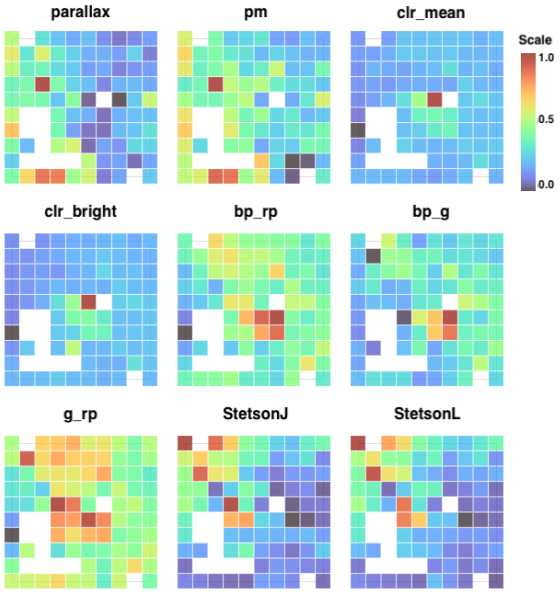


FIGURE 6.10: Feature maps for the XGBoost model produced using GTM. Same as for figures 6.9a and 6.9b though for Gaia and colour related features

contextual classifier was utilised to remove sources within the synonym radius (1.5") of a catalogued active galactic nucleus or nuclear transient. Inspection of light curves of alerting sources (within a 30-day period) removed under these conditions revealed no elimination of known or candidate CVs. To filter out supernova candidates, those sources classified as SN by Sherlock are removed should they meet the following criteria: the closest matching source from the PanSTARRS catalogue (used as the reference source) should have a Star/Galaxy score of less than 0.4 (values range from 0 to 1, where closer to 1 implies a higher likelihood of being a star); and an angular separation from the associated galaxy centre less than the galaxy's semi-major axis size (in arcseconds). Furthermore any source with a Transient Name Server name prefix with 'SN' was also removed. Of the sources remaining with a contextual classification of 'SN', $\sim 60\%$ displayed outbursting characteristics where quiescent stages were below the detection limit (likely dwarf novae). The remaining percentage was a mixture of faint sources with no star/galaxy score, several Mira variables, and a classified nova. For the removal of variable stars, a simple cross-match with the AAVSO VSX list of Mira variables, Cepheids, and RR Lyrae stars (amongst other classes under the variable star umbrella) was performed. None of those removed with this variable star filtering method belonged to a member of the confirmed or suspected CV family. Concerning the Δm criteria, no such filtering is performed to maximise the number of CVs. It was found that sources with the least amount of variability are assigned the nova-like class, thus a motivating factor

in this choice.

Constraining the number of alerts based on several g-r colour metrics, and not just the overall mean, had the desired effect of retaining dwarf nova exhibiting sources. These are outside the epochal or overall mean colour threshold of <= 0.7 during quiescence, but within the threshold during an outburst by virtue of the colour measured at their brightest epoch (*clr_bright*). An approximate quantitative estimate of the effectiveness of this strategy can be given for a month's worth of alerts. For June 2023, 12 confirmed or strong candidate dwarf novae were outside of this threshold based on the mean epochal or overall colour, whereas with the inclusion of the *clr_bright* quantity, only 1 fell outside the threshold.

An additional criterion requiring at least four data points for either the g or r band light curve was also imposed, allowing the majority of features to be derived. Combining all the above criteria, the number of sources returned per night for input into the ML classifier can be as few as 50, while on other nights over 200 may be available. During June 2023, the filtering output 1283 sources, of which $\sim$ 8% are contained within the Downes Catalog of CVs (Downes et al., 2001) and/or the Ritter Cataclysmic Binaries Catalog v7.24 (Ritter & Kolb, 2003). Approximately 45% are contained within the AAVSO VSX CV compilation of confirmed or candidate CVs (this includes the Ritter and Downes catalogues). The remainder, those not contained within AAVSO, comprise: low amplitude slowly varying (month to year-long timescales) sources ($\sim$ 30% of the total), a small fraction of which are eclipsing binaries; sources with similar variability to VY Scl and magnetic CVs ($\sim$ 3 and 4% of the total, respectively); outbursting candidates ($\sim$ 8% of the total); and a combination of sources that have once briefly risen above the limiting magnitude (possible supernovae), and those with too few data points for inference. Further inspection reveals that young stellar objects, candidate AGN, and variable stars provide the majority of contamination. A rough estimate of between 5 and 10% contamination from these sources is found.

The output of the filter applied to the alerts for June 2023 was fed into the XGBoost classifier with the following findings. The low variability sources are overwhelmingly assigned the nova-like class while outbursting sources are assigned one of the dwarf nova classes or the AM CVn label. Superoutbursting or candidate superoutbursting systems are largely assigned the SU UMa label with a small amount of mislabelling into the U

Gem class. Signatures of Z Cam variability are present within the list of sources assigned to this class, while faint blue sources are generally assigned to the AM CVn class. As one enters the low sampling regime (fewer than 20 data points) class confusion is evident, though not where outbursting activity is present.

From the June 2023 alerts filter output, I have compiled Table 6.7. This is a list of candidate CVs I identified that, at the time of writing, are not present in either the Ritter or Downes catalogues, the list of CVs within AAVSO VSX, or within the literature as far as I am aware. The prediction of class output by the classifier (along with the class probability) for these candidates is provided. Furthermore, I assign a score based on the strength of their candidacy as members of the CV class. A score of 1 represents a light curve sufficiently sampled for the identification of distinguishing characteristics. Should less well-sampled signatures of defining characteristics be present, for example, outbursts not sampled during quiescence, a score of 2 is given. A score of 3 is given to the examples where only faint signatures are present, possibly due to poor sampling.

TABLE 6.7: New CV candidates identified by my pipeline. Given are the: ZTF object ID; equatorial coordinates at the J2000 epoch; number of suspected dwarf nova outbursts, where (SO) is appended for possible superoutbursts amongst them; g band magnitude range, or r band (appended with r) should insufficient g band data exist (> is prepended should no quiescence brightness be present); light curve duration in days; Gaia BP-RP colour; mean ZTF g-r colour and in brackets, the colour at peak brightness, calculated in the manner of the *clr_mean* and *clr_bright* features explained in Table 6.3; prediction of our classifier; posterior class probability output by our classifier; and the strength of CV candidacy, rated as 1 for the strongest, 3 for the weakest candidates. The table is ordered by class prediction then probability.

ZTF ID	R.A.	Dec.	Outb	m Range	Dur	BP-RP	g-r	Clf pred	Prob	CV Rating
ZTF19aauxfaw	15:27:39.96	-19:48:46.17	4	> 17.9–19.1	1475	–	-0.34 (-0.17)	AM CVn	0.70	3
ZTF21aawqeix	18:49:31.03	-17:43:54.13	4	> 18.2–19.0	810	–	-0.02 (-0.08)	AM CVn	0.38	2
ZTF18ablpcfv	19:09:21.11	-20:01:03.13	6-8	> 17.5–18.7	1521	-0.60	0.03 (-0.08)	AM CVn	0.37	3
ZTF23aamdode	17:08:45.64	+08:54:51.69	1	> 17.4–20.5	44	–	-0.24 (-0.62)	AM CVn	0.35	3
ZTF19abdmfpn	17:58:04.69	+05:28:15.54	2	> 18.9–19.4	700	–	-0.44 (-0.27)	AM CVn	0.33	3
ZTF19aalcaij	18:01:43.65	+23:21:11.17	4-6	> 18.9–20.6	1409	–	-0.10 (-0.08)	AM CVn	0.31	2
ZTF19acbwtgi	22:25:56.91	+39:26:48.97	3	> 19.3–19.7	1375	–	-0.24 (-0.11)	AM CVn	0.30	3
ZTF18abcysck	19:03:59.30	+32:32:37.40	12 (SO)	> 18.5–19.7	1822	–	-0.33 (-0.31)	AM CVn	0.28	2
ZTF19aadovsk	17:44:08.17	-03:50:46.88	5-7 (SO)	> 18.5–19.3	1479	–	-0.16 (-0.04)	AM CVn	0.26	2
ZTF21acbqaqa	14:50:11.12	+65:59:42.19	–	18.9–20.7	654	–	0.30 (0.83)	Polar	1.00	2
ZTF20abpwtmi	15:38:20.42	+79:32:26.05	–	18.5–20.6	1071	–	0.38 (0.67)	Polar	0.96	2
ZTF18abcwxnq	18:43:26.49	+06:08:00.90	–	17.9–21.7	1153	1.86	0.27 (0.12)	Polar	0.94	2
ZTF18abmrmlu	23:01:52.75	+39:50:13.96	–	18.7–22.2	1791	0.91	0.41 (-0.12)	Polar	0.80	2
ZTF18abiklxf	20:46:40.96	+22:50:36.20	–	17.4–20.3	1816	1.46	0.22 (0.62)	Polar	0.77	2
ZTF18abnjsqz	17:40:39.30	-00:51:46.68	2	> 17.5–19.1	547	–	-0.04 (-0.01)	SU UMa	0.98	2

ZTF18abqbbpq	17:55:15.36	+06:57:44.41	4	> 18.6–19.9	1501	–	0.22 (-0.05)	SU UMa	0.98	3
ZTF19abtnbck	19:02:38.61	+26:52:44.76	3	> 18.8–19.7	1404	–	0.00 (0.00)	SU UMa	0.98	2
ZTF19abdolkk	19:21:46.43	-27:54:53.91	2	> 17.8–19.0	1454	–	-0.24 (-0.23)	SU UMa	0.98	2
ZTF19aaprbry	19:41:32.53	-07:37:54.12	4	> 18.6–20.0	1350	–	-0.05 (0.06)	SU UMa	0.98	3
ZTF20acufmrl	02:51:10.20	+48:39:28.83	3	18.5–19.9	263	–	-0.07 (-0.06)	SU UMa	0.97	2
ZTF19abjbhmd	16:55:20.72	-18:21:58.77	5	> 18.7–19.1	1442	–	-0.35 (-0.29)	SU UMa	0.97	3
ZTF19aalcaij	18:01:43.65	+23:21:11.17	1	> 18.9–20.6	1409	–	-0.10 (-0.08)	SU UMa	0.97	3
ZTF19aaxcajp	21:44:37.10	+29:30:10.74	5	> 18.3–19.7	1499	–	-0.11 (-0.02)	SU UMa	0.97	2
ZTF19aailtzw	17:07:44.19	+02:56:53.04	3	> 18.2–19.6	802	0.10	-0.09 (0.02)	SU UMa	0.94	2
ZTF18abcysck	19:03:59.30	+32:32:37.40	6	> 18.5–19.7	1822	–	-0.33 (-0.31)	SU UMa	0.93	2
ZTF21aaqwlgv	18:16:02.45	+03:07:11.79	3	> 18.3–19.5	819	–	0.05 (0.11)	SU UMa	0.92	2
ZTF18abklywy	18:01:53.06	+04:07:22.51	6	> 18.6–19.9	1526	–	0.23 (0.08)	SU UMa	0.91	2
ZTF19aadovsk	17:44:08.17	-03:50:46.88	3	> 18.5–19.3	1479	–	-0.16 (-0.04)	SU UMa	0.92	2
ZTF18aavtqlz	17:49:11.47	+23:58:27.57	5	> 19.2–20.3	1265	–	-0.24 (0.07)	SU UMa	0.85	3
ZTF18abthqde	19:39:04.33	+41:53:10.10	4	> 17.4–18.9	1760	–	-0.21 (-0.23)	SU UMa	0.83	2
ZTF20abylzfr	20:11:08.11	+84:05:19.21	2	> 17.1–19.7	1037	–	-0.08 (-0.16)	SU UMa	0.74	2
ZTF18absoqce	23:18:05.90	+55:58:51.90	6	> 17.9–19.4	1773	–	0.80 (0.39)	SU UMa	0.69	2
ZTF18ablpcfv	19:09:21.11	-20:01:03.13	5	> 17.5–18.7	1521	-0.60	0.03 (-0.08)	SU UMa	0.65	3
ZTF19ablvwcu	20:09:20.00	+00:22:28.56	5	> 17.7–18.5	1331	–	0.27 (0.19)	SU UMa	0.63	2
ZTF18abjrekr	22:00:29.91	+50:08:47.44	5	> 18.1–19.7	1808	–	0.20 (0.09)	SU UMa	0.62	2

ZTF18accpsgk	21:19:34.61	+38:00:12.90	10	> 17.2–18.0	1699	–	-1.30 (-0.85)	SU UMa	0.59	2
ZTF19ablujxj	20:36:53.40	+21:11:06.05	7	> 18.6–20.0	1438	–	-0.03 (0.00)	SU UMa	0.57	2
ZTF18abndsft	17:25:12.81	-20:40:48.85	4	17.7–21.2	1474	1.69	0.74 (0.53)	SU UMa	0.45	2
ZTF18abzmujj	19:11:51.25	-05:49:30.43	6	> 18.7–19.6	1730	–	0.62 (0.41)	U Gem	0.85	1
ZTF18abeajjd	17:03:58.75	+15:27:31.78	8	> 18.5–20.7	1823	–	0.13 (0.18)	U Gem	0.78	1
ZTF19aawxrtk	18:08:13.30	+22:51:09.39	2	16.9–17.2	1323	–	-1.73 (-1.42)	U Gem	0.68	2
ZTF18abloyve	19:10:41.97	-26:46:57.55	4	> 16.9–17.9	1490	–	0.44 (0.20)	U Gem	0.53	2
ZTF18aazeong	22:24:05.48	+51:11:42.41	10	17.3–19.3	1847	1.15	0.20 (0.11)	U Gem	0.47	1
ZTF18abnwfvw	18:53:33.53	+22:35:59.41	3	> 16.5–19.9	1422	1.64	0.54 (0.39)	Z Cam	0.45	2
ZTF18abuytrt	18:13:14.20	+01:49:02.04	> 9	18.2–20.8	1552	0.93	0.33 (0.40)	Z Cam	0.35	2
ZTF19aarpwtt	19:54:34.93	+46:11:08.59	10-14	> 18.8–19.8	1485	–	0.20 (0.08)	Z Cam	0.31	2
ZTF19ablujxj	20:36:53.40	+21:11:06.05	12 (SO)	18.6–20.0	1438	–	-0.03 (0.00)	Z Cam	0.31	2
ZTF18abthqde	19:39:04.33	+41:53:10.10	5	> 17.4–18.9	1760	–	-0.21 (-0.23)	Z Cam	0.30	1
ZTF21aaqwlgv	18:16:02.45	+03:07:11.79	3	> 18.3–19.5	819	–	0.05 (0.11)	Z Cam	0.25	2
ZTF18abnjsqz	17:40:39.30	-00:51:46.68	3	> 17.5–19.1	547	–	-0.04 (-0.01)	Z Cam	0.22	2
ZTF19aadospr	16:53:37.97	+00:49:11.93	4	> 18.4–19.7	805	0.36	-0.04 (-0.07)	Z Cam	0.21	3

6.4 Discussion

6.4.1 Classifier performance

The characteristics of the confusion matrix and the blending of class-specific structures into one another can be explained in the context of the physical properties of CVs, their evolution, and the properties of their light curves.

6.4.1.1 Class proportions

A list of thousands of cataclysmic variables accurately labelled into their subtypes based on multi-wavelength photometry with sufficient sampling and spectroscopy for each source is not currently available. While over 15,300 sources have been assigned the CV class according to the AAVSO and BTS, those with ZTF counterparts represented just over 5,700 (as of March 2023 when the dataset was constructed). A significant proportion of these belong to the dwarf nova class ($\sim 89\%$) of which only 19% possess labels with the dwarf nova subclass information required. I was therefore limited to a list of 1,439 sources with highly imbalanced class proportions.

Whilst efforts are made to account for this imbalance, the classes lowest in sample size (AM CVn, intermediate polar, and nova) are the weakest performers. Comparisons of light curves associated with each of these classes with the remaining classes provide a possible reason for their misclassifications. The intermediate polar ZTF17aabhicw (see Figure 6.2) displays long-term variability (weeks to months) as seen in polars, nova-likes and VY Scl (e.g., ZTF18abryuah and ZTF18abmrryp), while ZTF17aabglmw displays occasional dwarf nova outbursts. AM CVns display regular and super outbursts (e.g., ZTF18aaawjmk) and may be faint enough to only be visible during outburst (e.g., ZTF18adkhuxp), overlapping with SU UMa characteristics; longer-term changes associated with changes in mass-transfer rate (e.g., ZTF18aaabbbv) may also be present. A nova eruption decline (e.g., ZTF19aabjxpe) could be confused with SU UMa systems with long supercycles.

Despite these issues, the ROC curves and class maps represent a classifier with strong predictive capacity, even for the AM CVn and nova classes. This may be a consequence of features relevant to colour, parallax and proper motion. Nova systems in our sample

possess redder colours, while AM CVns typically lie at the blue end of the colour scale. AM CVns are intrinsically faint and, therefore, are required to be closer than most other CVs to be detectable and induce high values of parallax and, where tangential motion occurs, observable proper motion.

6.4.1.2 Dwarf nova classes

Distinguishing between different classes of dwarf novae primarily hinges on the features' ability to detect the presence of superoutbursts in SU UMa and standstills in Z Cam systems. In a study conducted by Otulakowska-Hypka et al. (2016), an in-depth analysis was undertaken to examine the characteristics of superoutbursts and normal outbursts in dwarf nova systems. The research revealed that Z Cam outbursts typically exhibit a noticeably lower amplitude range, spanning approximately 1–4 magnitudes, compared to the superoutbursts and normal outbursts observed in SU UMa systems, which range from 1–9 and 1–8 magnitudes, respectively. The upper limit for U Gem outbursts falls between these two extremes, with a range of 1–6 magnitudes. Consequently, one would anticipate significantly higher values for amplitude-related features for SU UMa compared to the Z Cam systems. Indeed, when examining the g and r band feature maps in Figures 6.9a and 6.9b for amplitude, the difference between the minimum (brightest) and mean or median magnitudes (*dif_min_mean* and *dif_min_median*), and the number of peaks with amplitudes exceeding 5 magnitudes (*npeaks_above5*), the highest values are consistently found within the region of GTM latent space occupied by SU UMa systems (see Figure 6.8 class maps). As we shift our focus from the SU UMa region in these class maps to U Gem and then to the Z Cam region, the feature values for the corresponding locations in the feature maps progressively diminish. The confusion matrix (Figure 6.5), along with those class maps, corroborates the notion that the most pronounced distinction among dwarf nova subtypes lies between SU UMa and Z Cam.

The semi-regular outbursts in dwarf nova systems exhibit a quasi-periodic pattern when adequately sampled. In ZTF light curves, it is notable that superoutbursts, especially long-lasting ones, tend to receive more comprehensive sampling compared to normal outbursts (refer to Figure 6.2). Consequently, the strength or amplitude of signals detected in the Lomb Scargle periodogram can serve as an effective discriminator for distinguishing SU UMa systems from U Gem and Z Cam. Notably, the feature maps within Figures

6.9a and 6.9b illustrate that the amplitude values corresponding to detected frequencies and their harmonics (referred to as *Freq**i**_harmonics_amplitude_**j***; see Table 6.2) are consistently higher in regions associated with SU UMa systems than in U Gem and Z Cam associated regions (refer to Figure 6.8 class maps). The peak values of these features are most prominent in regions adjacent to those associated with the AM CVn and nova classes, possibly due to instances where the observational timeline exclusively captures a brightening event, such as a nova eruption or superoutburst.

Figures 6.9a and 6.9b reveal that skewness (*Skew*), standard deviation (*Std*), and the standstill level (*stdstilllev_t20s10*), may be used to distinguish Z Cams from other dwarf novae. My analysis suggests that standstills can significantly influence the magnitude distribution, pushing it towards brighter values. Furthermore, if these standstills persist for an extended period, ranging from weeks to months, they can also reduce the standard deviation, aligning it more closely with that observed in nova-like systems. While regions exhibiting low standard deviation are not exclusive to Z Cam systems, as other dwarf novae with extended periods of quiescence also display this characteristic, what sets Z Cams apart is the normalised brightness within these low standard deviation regions. The standstill level feature aims to pinpoint these distinctive regions within the light curve, effectively distinguishing Z Cam systems from their SU UMa and U Gem counterparts.

When it comes to defining characteristics of U Gem systems, with orbital periods greater than 3 hours, their more massive donor stars and greater mass-transfer rates result in accretion disks typically larger than those of SU UMa systems, whose orbital periods mostly lie below 2 hours. Consequently, for the equivalent orbital inclinations, U Gem systems have a higher optical quiescent brightness. The combination of ZTF's limiting magnitude and this brightness disparity results in many SU UMa systems only being detected during their outburst phases as opposed to the U Gem class in which quiescence sampling is more likely. This is evident when examining the number of observations (*n_obs*) feature maps in Figures 6.9a and 6.9b, where higher values are present in the U Gem associated region compared to that for SU UMa.

Expanding upon the topic of intrinsic brightness, sources with lower intrinsic brightness would need to be closer for effective observation, leading to a higher parallax measurement (and possibly proper motion depending on motion in the tangential plane). With

the shortest orbital periods of the dwarf nova classes, SU UMa systems are expected to be less luminous, (given equivalent orbital inclinations) for the reasons set out in the previous paragraph, and possess higher parallax values (and proper motion) when compared to their dwarf nova counterparts. These distinctions are indeed evident in the Figure 6.10 feature maps for *parallax* and *pm*, respectively. Moreover, these arguments align with the observation of fainter absolute magnitudes as well.

The high mass-transfer rates characteristic of Z Cam systems drive them to meet the disk instability threshold shortly after a previous outburst. Consequently, during their outburst phases, they tend to spend considerably less time at the minimum brightness level in comparison to other dwarf nova types, as documented by Simonsen et al. (2014). This leads to recurrence periods typically falling within the range of 10 to 30 days, exemplified by systems like ZTF17aaaeepz. It is reasonable to anticipate that the outburst recurrence period, a parameter that the Lomb Scargle periodogram's maximum power frequency (*freq_pwr_max*) aims to characterise, could offer some level of discrimination between Z Cam systems and their dwarf nova counterparts.

However, upon scrutinising the corresponding feature maps for *freq_pwr_max* (within Figures 6.9a and 6.9b), it becomes evident that distinguishing between these types is challenging. For potential insights into this challenge, one may refer to the findings of Otulakowska-Hypka et al. (2016). Notably, while the average recurrence periods for the U Gem class tend to be longer than those of Z Cam systems, over 50 days, there is an overlapping range with Z Cam recurrence periods. This overlap is also observed in the case of the SU UMa class, where recurrence periods span from 3 to 300 days. Additionally, factors such as the presence of extended standstills in Z Cam systems (e.g., ZTF17aabunpt; Figure 6.2) and the limited sampling of normal outbursts contribute to the complexity of estimating this type of periodicity.

An examination of light curves for systems that fall between the latent space nodes associated with U Gem and Z Cam classes (see Figure 6.8) further confirms this recurrence period overlap, as does the overlap between the SU UMa and U Gem classes. Within this continuum also lie the rapidly outbursting SU UMa subtypes, ER UMa, underscoring the significance of recurrence period overlap as a primary contributor to the confusion among dwarf nova subclasses.

6.4.1.3 AM CVn

For the remainder of Section 6.4.1, to facilitate our discussion and interpretation of the class and feature maps, I may refer to specific nodes (squares) by a simple coordinate system (x, y). The value of x denotes the square number (1–10) from left to right, while the value of y signifies the square number (1–10) from bottom to top.

As previously discussed in the introduction, AM CVn systems tend to be bluer than their hydrogen-rich CV counterparts and are generally of lower luminosity. While superoutbursts are observed in AM CVn systems (Kato & Kojiguchi, 2021), they tend to be of shorter duration, typically lasting 5–6 days, and display lower amplitude (4–6 magnitudes) in contrast to superoutbursts in SU UMa systems, which often extend beyond 10 days and can, in the case of the WZ Sge subclass of SU UMa, reach amplitudes exceeding 6 magnitudes. Additionally, normal outbursts have also been observed in AM CVn systems, occurring on the fading tail of superoutbursts (Duffy et al., 2021).

Upon scrutiny of feature maps, it becomes apparent that features such as the mean, median, minimum, and maximum magnitude derived from g-band light curves (Figure 6.9a) do not strongly differentiate AM CVn systems from other classes, contrary to the expectation of higher (and consequently fainter) values. Similar observations hold for the r-band (Figure 6.9b), except for the minimum magnitude in the r-band (min_mag_r), where notably elevated (i.e., fainter) values cluster around node (1,4), associated with the highest AM CVn probability (see Figure 6.8 class maps). One possible explanation for these findings is that accretion discs in AM CVns are smaller than those in hydrogen CVs, truncated by the smaller Roche lobe geometry. As emissions in the r-band primarily originate from the cooler outer regions of the accretion disc, the effective surface area of these regions is considerably smaller for the compact AM CVn discs.

To become detectable, AM CVn systems would be required to be situated at closer distances, thereby inducing higher parallax measurements and, in cases where tangential motion occurs, observable proper motion (pm). While node (1,4) within the corresponding feature maps in Figure 6.10 may not contain the highest values (which are located at node (3,7) and associated with the SU UMa region), they still exhibit values sufficiently high enough to align with expectations when compared to regions associated with other classes.

The average ZTF g-r colours, along with Gaia colours (involving RP data), are strong discriminators effectively separating AM CVn systems from other classes, as evident in Figure 6.10. However, when it comes to outburst-specific features (e.g., *npeaks_2to5*; Figures 6.9a and 6.9b), their effectiveness diminishes. Contributing factors to this reduced performance may the scarcity of AM CVn examples within the dataset, coupled with variations in observational time-spans and the sampling of their light curves. Consequently, this diversity results in a variety of light curve profiles, as depicted in Figure 6.2, where the number of sampled outbursts ranges from several to none at all. An examination of sources projected onto latent space regions where the boundaries between AM CVn and SU UMa classes, as well as between AM CVn and nova classes, blend (see Figure 6.8), suggests that these factors contribute significantly to the observed classification ambiguity.

6.4.1.4 Novae

Despite a low sample size, the nova class achieves a recall score of 0.50 and a precision of 0.64. A significant source of false-positive predictions in the nova class can be attributed to the SU UMa class. A possible explanation could simply be due to nova dataset examples consisting largely of extragalactic sources, visible during the time of peak eruption brightness. These light curves bear a resemblance to those of SU UMa systems where only one outburst (often a superoutburst) has been sampled. Consequently, a low number of observations is associated with the class, as is the case for SU UMa systems.

Two members of the nova class within the test set have misclassifications as VY Scl. A possible explanation could be provided by ZTF21abmbzax (example light curve in Figure 6.2), which displays a 'dust dip' explained as being generated by dust in the eruption ejecta absorbing photons and re-emitting in the infra-red (Strope et al., 2010). This characteristic resembles a VY Scl low-state excursion. Another eruption light curve profile mentioned in Strope et al. (2010) exhibits a 'flat top and jitters' — cuspy profiles at eruption maximum. This is seen in ZTF19abirmkt and could be responsible for misclassifications of novae as magnetic CV members. Projections of these sources onto the GTM latent space of Figure 6.8 align with these interpretations, with ZTF21abmbzax projected onto node (6,3), located in between the nova and VY Scl structure cores, and

ZTF19abirmkt projected onto node (3,3) located between the nova and polar structure cores.

6.4.1.5 Remaining classes

The separation between the intermediate polars, polars, nova-likes, and the VY Scl nova-like subtype arises from several physical properties manifested in their light curves, as discussed in the introduction. As just demonstrated in previous subsections, a comparison of the g and r band feature maps within Figures 6.9a and 6.9b with the class probabilities depicted in the Figure 6.8 class maps, help highlight the light curve attributes most relevant for class separation.

The VY Scl class stands out with its deep low brightness state excursions such that low values of *eta_e* appear in the relevant g and r band feature maps near node (7,1), associated with the highest VY Scl class probability (see class maps). This feature reflects the degree of independence between successive data points, where magnetic systems exhibit higher values due to hourly timescale variations, while VY Scl systems show lower values due to longer timescale variations. Furthermore, VY Scl low-state excursions can induce a high skewness in magnitudes (*Skew*), and due to stable and prolonged high-brightness states, give rise to the highest standstill level values (*stdstilllev_t20s10*), as feature and class maps demonstrate.

Eclipses within the nova-like class, as exemplified by ZTF18abajshu in Figure 6.2, push the standstill level into a range occupied by Z Cams, potentially causing confusion between these two classes. Confusion also arises between nova-likes and the SU UMa class. The light curves of sources where such confusion occurs are marked by a limited number of data points, this is seen in the *n_obs* feature maps for nodes (6,6) and (6,7), situated where the associated class structures are closest together. Based on the current feature set, the model finds difficulty in distinguishing systems visible only during outbursts from nova-likes with limited observational epochs, though overall, nova-likes remain distinguishable from the other classes.

The lowest standard deviation (*Std*) and median absolute deviation (*MedianAbsDev*) values are associated with the intermediate polar and nova-like classes, as seen in the

feature maps. This aligns with the less frequent low-state excursions observed in intermediate polars and nova-likes compared to polars and VY Scl systems in the ZTF light curves.

As explained by Hameury & Lasota (2017), most intermediate polars possess accretion disks truncated at inner radii due to the white dwarf's magnetism. This may lead to dwarf nova outbursts characterised by lower amplitudes and shorter durations. The mixture of outbursting and non-outbursting intermediate polars, coupled with less distinct outburst profiles, contributes to feature maps displaying lower amplitude and variability-related values for intermediate polars compared to dwarf novae. Non-outbursting intermediate polars may explain the confusion with polars, indeed, this is supported by the projection of intermediate polar ZTF18abaiuvj (Figure 6.2) onto a region associated with polars within the GTM latent space (Figure 6.8).

6.4.1.6 Evolutionary factors

Separating cataclysmic variables into distinct classes is one that is a challenge for experts on the subject who must wrestle with the fact that as these systems evolve, they transition from displaying traits characteristic of one class to another such that boundaries between classes are blurred (e.g., Warner 1995; Hellier 2001; Förster et al. 2021; Paczyński 1971; Shafter 1992).

The shortening of orbital periods, donor composition changes, and shrinkage of the accretion disk amongst several other factors drive the class transitions. Nova-likes have mass-transfer rates high enough to be stable against dwarf nova outbursts. As the mass-transfer rates drop, the accretion disk straddles the stability threshold, below which the disk is cool, non-viscous and unstable to dwarf nova outbursts (Shafter, 1992). The Z Cams, which lie close to this threshold, provide a link between non-outbursting and regularly outbursting dwarf novae, with periods of standstill, akin to nova-likes, and outbursting episodes typical of dwarf novae. The continuing evolution induces an unstable disk, where a transition to the semi-regular outbursts of U Gems and then the superoutbursting SU UMa systems occurs. The ER UMa subtype of SU UMa dwarf novae may cause confusion with the Z Cam class, where above-average mass-transfer rates lead to short superoutburst recurrence periods and rapid-fire normal outbursts. As with hydrogen CVs, helium CVs (AM CVns) undergo an evolution (to longer periods)

that results in observational changes, during which accretion may transition from direct (no disk), to hot stable, then unstable disks subject to the He CV equivalent of dwarf nova outbursts (Nelemans, 2005; Solheim, 2010).

To add to the difficulty in assigning class labels is the presence of nova eruptions — a possibility for all systems should conditions for hydrogen fusion be present under degenerate conditions on the WD surface; this is far more likely to occur for the highest mass-transfer rate systems with high mass WD accretors (e.g., Munari 2012; Chomiuk et al. 2020; Darnley et al. 2006; Darnley & Henze 2020). The presence of strong magnetic fields for the intermediate polar or polar label requires observations of pulsed X-rays and/or polarimetry to complicate matters further. In addition to the above, one must factor in the orbital inclination that determines the overall emission contribution from the accretion disk and thereby impacts measurements such as colour and brightness.

The evolutionary changes are evident in many of the light curves in my dataset. CR Boo (ZTF18adkhuxp; Figure 6.2), is an AM CVn with a standstill to its name (Kato et al., 2023); high mass-transfer rate systems residing amongst the U Gem class manifest as dwarf nova outbursts with very short recurrence times indicative of a Z Cam class; the ER UMa subclass (Kato et al., 2013) of SU UMa systems may also be confused with the Z Cam class due to their high mass-transfer rates and rapid outbursts. Concerning intermediate polars a range of light curve morphologies are possible (e.g., Šimon 2021). Short duration low state transitions, dwarf nova outbursts and more stable long-term light curves are present within our light curve sample, consequently, confusion with any of the other classes is possible. Constructing a classifier in light of these intricacies will naturally produce class confusion despite incorporating a wide-ranging feature set inclusive of astrometric data and an attempt to produce a dataset with accurate class labels. The confusion matrix and class maps in combination with the example light curves displayed in Figure 6.2 are a visual representation of this very aspect of CV classification.

6.4.2 Pipeline implementation

A substantial portion of the alert stream filter consists of either known or candidate CVs (according to the AAVSO VSX list). This is positive news, indicating that the

filter effectively retains them in the stream. Consequently, the undiscovered CV candidates have a promising likelihood of being contained within the remaining alerts that have successfully passed through the filter. The approximate class proportions of the confirmed or candidate objects are as follows: 20% SU UMa (including the WZ Sge and ER UMa subtypes), 4% Z Cam, 3% U Gem, 59% dwarf novae without further subdivision, 3% magnetic CVs, 6% nova-likes (including subclasses), and less than 1% AM CVn. The remaining confirmed or candidate CVs form a mixture of several sources labelled as novae due to an eruption that may have occurred before ZTF observations, a recurrent nova, and CVs without further subdivision. These proportions stem from a variety of factors that may include: the frequent occurrence of alert-triggering events in dwarf novae leading to their relatively higher representation; the inherent faintness of short (or ultrashort) period CVs, making their detection less probable; the need for supporting evidence, such as periodic variability on short timescales (minutes to hours), polarimetry, and/or X-ray emission, to confidently confirm a CV as magnetic; and the establishment of specific thresholds in the alert filtering, e.g., excluding CVs with a g-r colour index exceeding 0.7.

The substantial contribution of low-variability sources among the remaining filter targets results from the omission of a magnitude change condition. Nevertheless, it was observed that incorporating such a condition restricted the detection of confirmed/candidate outbursting CVs unless a considerably low threshold was applied. Notably, our classifier overwhelmingly assigns the nova-like label to low (or slowly varying) sources, thereby enabling the classifier to allocate the remaining higher variability sources into distinct classes. Nonetheless, we retain the option to implement a magnitude change criteria should we choose to focus on specific variability types.

Referring to Figure 6.1, configuring the filter to retain alerting sources with a ZTF g-r colour of $<= 0.7$ is expected to encompass the vast majority of the shortest period systems, SU UMa, and AM CVn candidates, along with a significant portion of the remaining classes. However, expanding the filter to include all examples would inevitably lead to a rise in contamination from non-CVs, such as Mira variables and AGN candidates (as observed in the June 2023 sample). Similar to the magnitude change filtering, I am actively exploring the option to adjust the colour constraint, aiming to focus on specific CV subclasses.

The ML classifier demonstrates its greatest strength when applied to the filter output by effectively distinguishing between outbursting and non-outbursting sources, a characteristic mirrored in the test set predictions. Also mirroring the test set results is the further separation of confirmed, candidate, or likely (from inspection) SU UMa from Z Cam sources; and the separation of light curves with polar and VY Scl-like variability assigned to those respective classes. However, when we enter the low sampling regime, the classifier struggles to assign alerting sources into what I would consider the appropriate class. For example, several poorly sampled though likely outbursting systems (where quiescent magnitudes are not sampled) are assigned the nova-like or polar classes. However, on the whole, these sources tend to be assigned one the dwarf nova classes or the AM CVn class (should an especially blue colour be calculated).

6.5 Conclusions

In this paper, I developed and applied a machine learning pipeline to detect and categorise cataclysmic variables (CVs) and their subtypes from the ZTF alerts stream. The pipeline's alert filtering stage effectively retains both known and potential CVs across various subclasses, thanks to a multi-parameter g-r colour threshold and the omission of a magnitude change condition. This approach accommodates colour changes during dwarf nova outbursts.

The performance of the ML classifier is largely dependent on the ability of the dataset to provide an accurate representation of the diversity within the CV population. This diversity is present in the example light curves (see Figure 6.2), however, imbalance in this diversity (class imbalance) and commonalities in the types of photometric variability between classes renders CV subtype classification a particularly challenging task. Evolutionary factors drive the difficulty in arriving at concrete class labels both for experts in the subject and the ML classifier. The challenge is compounded by inadequate sampling of light curves. Despite these difficulties, an exhaustive examination of several ML algorithms, trained with a comprehensive feature set, and operating under a selection of class balancing and feature selection techniques, yielded a classifier with a prediction pattern that can be understood in the context of CV evolution.

Latent space representations of this prediction pattern using GTM (class maps) provide an easily interpretable avenue for visualising this evolution. The accompanying feature maps provide a convenient method of finding those features most relevant for a model's assignment of a given class. They also provide us with the properties that contribute to classification error, where in many cases the answers are linked to evolutionary factors. Though not explored in this work, these feature maps provide a method to pare down the feature set by eliminating features that provide little benefit for discrimination between classes.

Implementation of the pipeline on the ZTF stream has, over the period of June 2023 alone, yielded a sample of 51 new CV candidates, These are largely outbursting, with several magnetic CV candidates. With further improvements to the pipeline underway, such as filter threshold adjustments and the inclusion of computer vision techniques to provide an automated interpretation of salient light curve characteristics, I aim to reduce contamination of non-CVs (e.g, Mira variables and active galactic nuclei) and produce an ML classifier with greater class distinction powers.

Given the fuzzy boundary between CV subclasses for the reasons mentioned, it may be prudent to apply stricter criteria for dataset inclusion, focusing only on clear examples of a given class. With this approach, one relies less on definitive class labels, but more on the probability of class belonging. Alternative approaches may include adopting a multi-label approach that takes into consideration class boundary crossing variability, or an unsupervised learning strategy that does away with existing class labels, tasking algorithms with finding similarities, differences and structure in the data itself.

Chapter 7

Unsupervised Learning

7.1 Introduction

In the field of machine learning classification, label noise (or incorrect class labels) impacts the training of models and can lead to suboptimal classification performance. Label noise may arise due to many reasons, for example, human error, insufficient/incomplete data, ambiguity in defining class boundaries, examples with diverse and complex behaviour; or examples exhibiting characteristics of multiple classes. Adopting an unsupervised learning strategy can help to address these issues. Dimensionality reduction is especially useful in this regard. By projecting the data into a lower dimensional space the intrinsic structure of the data can be revealed enabling a comparison with our preconceived classification structure. This intrinsic structure may then be further analysed to understand the similarities and differences of properties (features) between any groups of examples present. Clustering algorithms may also be useful in defining a classification structure based on the data alone which can be compared to existing class labels. However, fuzzy class boundaries can make it difficult to identify the appropriate number of clusters.

7.1.1 Unsupervised Learning Examples in Time Domain Astronomy

Within the field of time-domain astronomy, unsupervised learning has been applied for examining the validity or substructure of classification schemes, assessing the usefulness of features, and as an initial step in a classification pipeline that can identify particular

classes/groups of time-varying objects from survey data. Several recent examples exist. Particularly relevant to this research is the search for CVs within Data Release 6 (DR6) of the LAMOST survey containing nearly 10 million low-resolution spectra (Sun et al., 2021). The search revolved around their classification based on the presence of Hα emission lines. The process involved a training set comprising a set of 392 LAMOST CV (or candidate CV) spectra showing Hα emission and 973 non-CV LAMOST spectra without Hα emission. The flux measurements in the Hα region (6530–6600Å) were used as input for UMAP to reduce the input to 3 dimensions. The UMAP output is used as input for the K Nearest Neighbours algorithm to output a classifier tasked with distinguishing between Hα emission spectra and spectra without Hα emission. Implemented on the DR6 data, 169,509 of the $\sim$10 million spectra were identified as possessing Hα emission. From here a combination of spectral model fitting and manual inspection resulted in 323 CVs or candidate CVs being identified, of which 52 are new candidates. The utilisation of UMAP helped considerably in paring down a large dataset.

Another example is related to the field of Gamma-Ray Bursts (GRBs). The traditional picture of long GRBs (>2s) originating from core-collapse supernovae (Woosley & Bloom, 2006; Hjorth & Bloom, 2012) and short GRBs from compact binary mergers involving a neutron star (NS) leading to kilonovae (Paczynski, 1986; Goldstein et al., 2017) has been challenged by recent observations (Rossi et al., 2022; Lü et al., 2022), requiring a more detailed classification structure. Dimple et al. (2023) chose to examine this problem by using Swift XRT light curves as input for PCA initialised UMAP and t-SNE algorithms. They subsequently employed AutoGMM, a clustering algorithm that utilises a Gaussian mixture model to represent the data as a combination of Gaussian distributions. AutoGMM is employed to determine the optimal number of clusters within the dataset. Dimple et al. (2023) found that five distinct clusters of GRBs exist, of which the kilonova-associated GRBs are located in two separate clusters. Their use of unsupervised learning led to an interpretation that these may be due to different progenitors — subclasses of binary neutron star and/or NS–black hole mergers.

Narayan et al. (2018) employed PCA initialised t-SNE to examine the usefulness of a set of statistical features extracted from OGLE and the Open Supernova Catalog light curves. They aimed to distinguish between several variable star classes and supernovae. The t-SNE representation divided large sample size classes (eclipsing binaries and RR

Lyrae) into subclusters, which for RR Lyrae turned out to be a reflection of their subclasses, while Cepheid variable classes were fairly well separated. However, much class overlap remained and low sample size classes were not well distinguished. Narayan et al. (2018) re-implemented the procedure, this time with a dataset balanced with a combination of random undersampling and oversampling with a SMOTE variant. Narayan et al. (2018) used the clearer separation in the resultant projection as a visual representation of the adverse effects of class imbalance within their data.

The ZTF source classification project detailed in van Roestel et al. (2021) describes a hierarchical ML pipeline for the classification of transient/variable star sources detected by the ZTF transient alert stream. The training set classes included YSOs, different classes of variable stars, and AGN. The application of t-SNE was a means to reveal the dataset structure based on their feature set. While the separation of variable sources from non-variable ones is clear, the separation of the aforementioned classes revealed a representation with a complicated substructure and class overlap (especially for classes with fewer examples).

7.1.2 Unsupervised Learning for Cataclysmic Variables

The above are examples of how one may implement dimensionality reduction for exploring classification structures, aiding in classification, and examining feature relevance. However, to the best of my knowledge, no research explores the diverse range of variability within the CV family with unsupervised learning methods. While Sun et al. (2021) used UMAP to help identify CVs from the LAMOST spectroscopic survey, further examination of the Hα emission spectra for substructure was not explored. Feature validity for several broad transient classes was explored by Narayan et al. (2018), though this has yet to be explored for CV subclasses. In this chapter, I address these gaps in the CV literature. Here, dimensionality reduction is explored in an attempt to: elucidate the true diversity of examples within the CV dataset; assess the relevance of the features used in the classification models; and identify new members of particular CV subclasses from a list of sources with only a broad classification. So far I have cited instrumental and evolutionary factors as a cause for the challenges faced when adopting the existing classification scheme. Explicitly, inconsistent or sparse light curve sampling, telescope limiting magnitude, label noise, CVs transitioning between classes (or similar

evolutionary factors), class imbalance, and defining class characteristics not always being present during the observational timespan. I provide a more detailed examination of the challenges faced through the examination of outputs from dimensionality reduction algorithms PCA, t-SNE, UMAP, and GTM.

PCA is concerned only with linear separability within the data, identifying the principal components (dimensions) that account for the greatest amount of variance in the data. PCA tends to preserve the global structure of high-dimensional space in its low-dimensional projection at the cost of local structure. For non-linear methods, t-SNE is focused on pairwise distances in data space, thereby preserving local structure at the expense of global structure. UMAP represents a middle ground between the global structure-focused PCA and the local structure-centric t-SNE. A major advantage of using UMAP over t-SNE is that UMAP results in a model for inference. A probabilistic model of the higher dimensional distribution of examples can be constructed using GTM. Through associated reference maps, one may visualise this probability distribution and identify where in feature space our clusters lie. They provide a method for comprehensively understanding the data distribution allowing the properties of similarly grouped CVs to be easily identified. With this selection of algorithms, I aim to cover the bases of linear and non-linear representations, global and local structure preservation, and interpretability of clusters.

The structure of this chapter is as follows. Section 7.2 covers the construction of the dataset, data preprocessing, the algorithm optimisation procedures and the tasks explored. Section 7.3 displays the resultant projections and their analyses with the aid of feature projections and reference maps. In section 7.3.5, I project example CVs not used during the training of the PCA, UMAP and GTM models, onto the low-dimensional space learnt by UMAP. These CVs have not previously been assigned subclass labels, therefore, their locations on the low-dimensional space provide an update to their present classification status whilst also providing an assessment of the generalisation ability of the models. The chapter concludes with Section 7.4, where the findings are discussed in the wider context of the research into CVs.

CV class	Count
Dwarf nova: SU UMa	378
Dwarf nova: WZ Sge	32
Dwarf nova: ER Uma	25
Dwarf nova: U Gem	115
Dwarf nova: Z Cam	168
Nova-like	138
Nova-like (VY Scl)	117
Polar	110
Intermediate polar	49
AM CVn	35

TABLE 7.1: A breakdown of the classes of CV present with the dataset for unsupervised learning analysis.

7.2 Method

7.2.1 Dataset construction

The features derived from the light curves, as described in Chapter 6, each require a minimum number of data points to be calculable otherwise null values are recorded. Furthermore, having more data points allows for a more detailed and accurate characterisation of the light curve. Therefore, a minimum data points threshold is set for an example to be included for dimensionality reduction analysis. In addition to this, the impact of the addition of Gaia DR3 data is also assessed. Therefore, for each algorithm, 2D projections under the following conditions were obtained:

- 1) Minimum data points threshold of 20 in either the g or r band and without the use of Gaia DR3 data;
- 2) Minimum data points threshold of 20 in either the g or r band with the use of Gaia DR3 data.

While experimentation with stricter thresholds (above 20 points) and applied to both rather than either filter band was conducted, any difference between projections was minimal to none. Also, one reduces the number of examples under consideration. Therefore, only the above conditions are explored in this chapter. For those conditions, Table 7.1 provides the number of examples for each class of CV.

7.2.2 Data preprocessing

The following strategy has been adopted to handle features with outliers, missing data, and heavily skewed distributions. Each feature is inspected for outliers to determine if they are erroneous, such as data points falling outside the accepted range. I handle outliers by capping feature values based on their correct range or using the interquartile range method. In the interquartile range (IQR) method, values above $Q3 + q \times \text{IQR}$ or below $Q1 - q \times \text{IQR}$ are capped at these boundary values. Here, $Q1$ (the first quartile) marks the 25th percentile, $Q3$ (the third quartile) marks the 75th percentile, and q is assigned a value between 1.5 to 3 (inclusive), depending on the feature distribution. Missing data in light curve-derived features is addressed using the approach outlined in Section 6.2.9. Specifically, missing values are imputed using feature values derived from the light curve of the other filter or a closely related feature (e.g., the overall mean $g-r$ colour if the epochal mean $g-r$ colour is unavailable). Any remaining missing values are handled using the K-Nearest Neighbors (KNN) imputation method, with the number of neighbours set to 5. Heavily skewed distributions can adversely affect the outcome of dimensionality reduction algorithms, for example, PCA is sensitive to the scale of features such that principal components are influenced by the tail of the distribution possibly leading to biased representations. Handling heavily skewed features with log transform can help to mitigate these effects and help with feature scaling, therefore this is the approach adopted.

7.2.3 Hyperparameter Optimisation

The quality and interpretability of the reduced-dimensional representations depend on the choice of hyperparameters, as small changes can lead to significantly different outcomes. Adjusting the hyperparameters allows us to balance the representation of local versus global structure in the data, prevent overfitting, and identify the optimal range where small adjustments or different random initializations do not lead to drastically different results. A reminder of the associated hyperparameters of the dimensionality reduction algorithms is summarised here.

- PCA:

  - *n_components* sets the number of dimensions in the low-dimensional representation. Set to 2.
  - no other adjustments

- t-SNE:
  - *perplexity* effectively sets the number of nearest neighbours each point is attracted to, where the larger (smaller) the value, the more global (local) structure will be present in the projection.
  - *early_exaggeration* controls the initial exaggeration of similarities between data points, helping to make clusters better separated during the early stages of optimisation.
  - *learning_rate* controls the step size of the gradient descent function as applied to the KL divergence loss function. Adjustments help to avoid getting stuck in a bad local minimum.
  - *n_iter_without_progress* sets the maximum number of iterations without reduction in the loss function before optimisation is aborted.
  - *metric* controls the method used to measure point-to-point distances in data space. Set to default of 'Euclidean'.

- UMAP:
  - *n_neighbours* sets size of local neighbourhood (in terms of the number of neighbouring sample points) used for manifold approximation. Larger values = more global structure emphasis, smaller values = more local structure emphasis.
  - *min_dist* sets the minimum distance apart that points are allowed to be in the low dimensional representation. Lower values result in clumpier embeddings useful for fine topological structure analysis, while large values are useful for broad structure analysis.
  - *learning_rate* controls the optimisation step size.
  - *metric* is as defined for t-SNE, and set to 'euclidean'.

- GTM:
  - k sets the square root of the number of GTM nodes

- m is the square root of the number of radial basis function (RBF) centres that approximates the data space probability distribution
- s sets the RBF width factor to tune the width of the RBF functions.
- *regul* is the regularisation coefficient.

7.3 Results and analysis

For each algorithm, projections generated with and without Gaia DR3 data features are shown, each colour-coded by class and feature values. To be more specific, the first figure for each algorithm presents projections (subplots) generated without Gaia DR3 data features. Each subplot displays data points colour-coded based on whether they belong to a specific CV class (one versus rest), with the final subplot colour-coded in a multi-class format. The second figure follows the same layout but shows projections generated with the inclusion of Gaia DR3 data features. These plots are referred to as 'class projections.'

The third figure for each algorithm displays the same projections (subplots) as the first (generated without Gaia DR3 data features) but with the data points colour-coded by their value for a particular feature (scaled between 0 and 1). One subplot is provided for each of a selection of non-Gaia DR3 features. The fourth figure uses the same layout as the third, but the projections are generated using Gaia DR3 data features, and subsequently, subplots are provided also for Gaia DR3 data features. The third and fourth figures will be referred to as 'feature projections.'

For GTM, instead of feature projections, reference maps are shown, which will be explained later. These maps offer a clearer method than the feature projections for elucidating the properties of clusters within the low-dimensional space.

Finally, projections of example CVs that were not used during the training of the PCA, UMAP, and GTM models are shown on the low-dimensional space learnt by UMAP (both with and without Gaia DR3 data features) to update their current classification status. These CVs have either not been assigned a CV subclass label or have been classified as a dwarf nova without a specific dwarf nova subtype label.

7.3.1 PCA

FIGURE 7.1: PCA 2D projection of dataset where a minimum points threshold of 20 in either the g or r band was set and no external (DR3) data was utilised. They are colour-coded by class, and presented in a one-versus-rest manner apart from the plot on the bottom right, which combines the preceding plots.

FIGURE 7.2: Same as Figure 7.1, though with the inclusion of external data from Gaia DR3

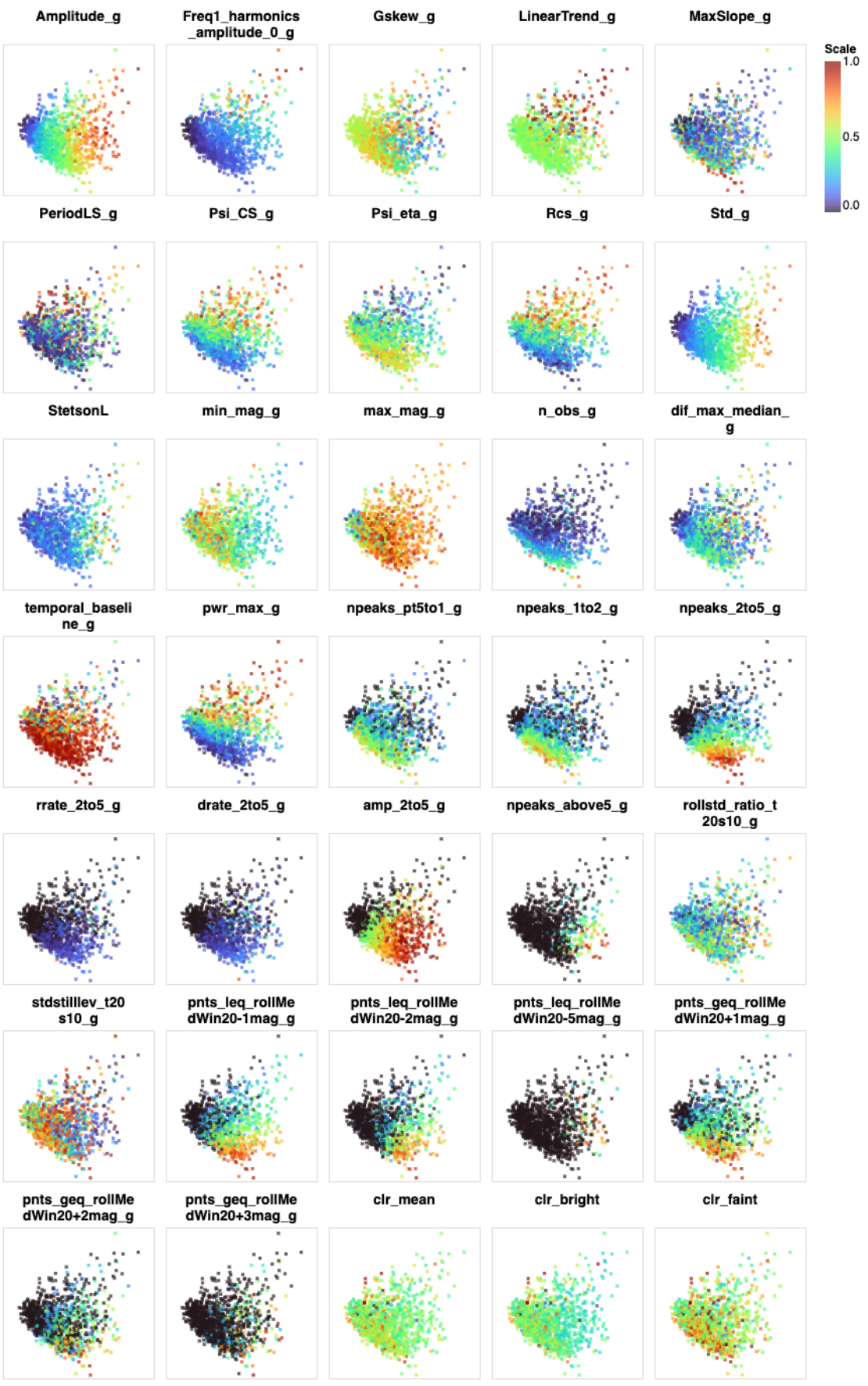


FIGURE 7.3: PCA projections without inclusion of Gaia DR3 data (see Figure 7.1) colour coded by feature values for selected features and scales to between 0 and 1.

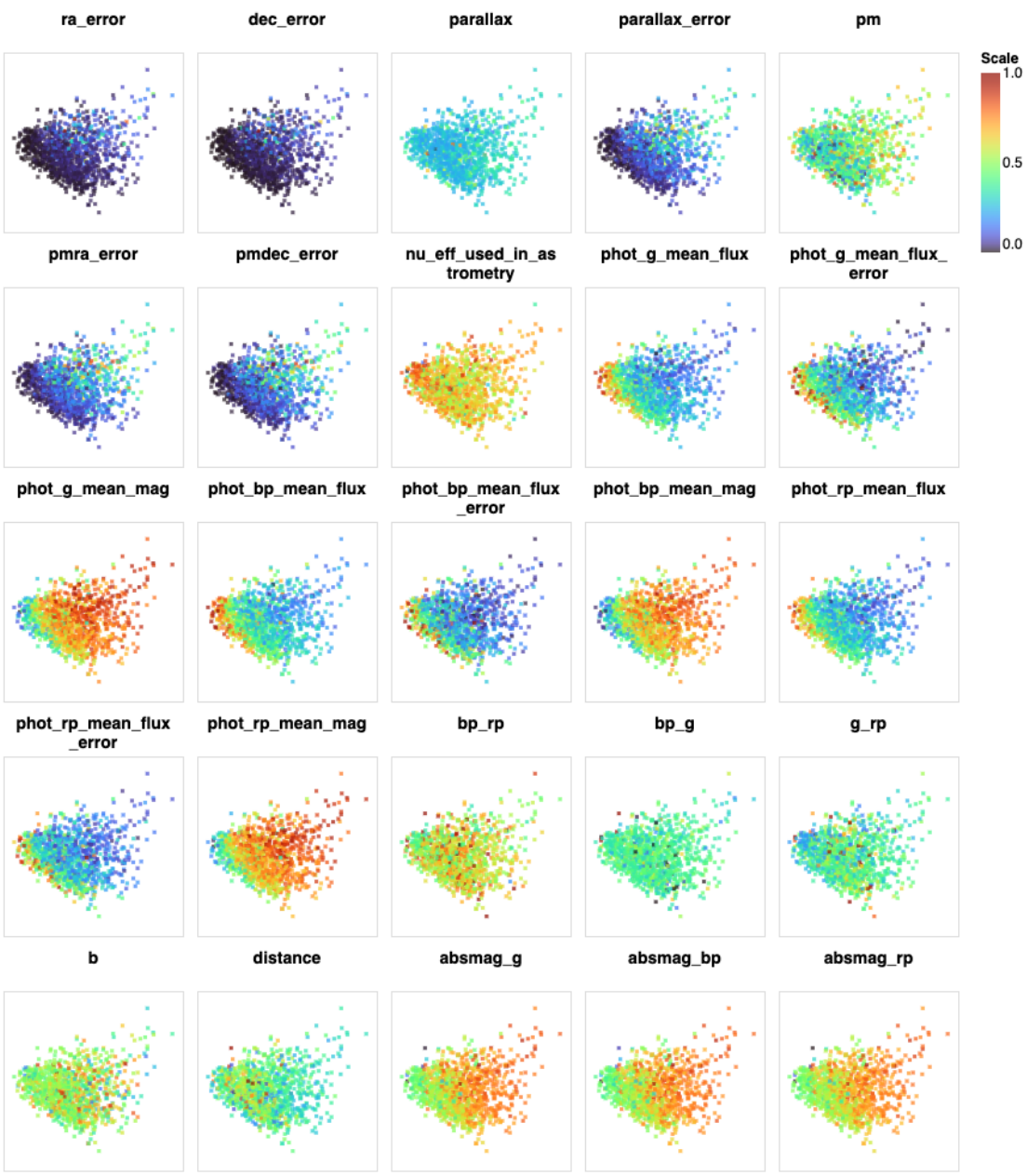


FIGURE 7.4: PCA projections with inclusion of Gaia DR3 data (Figure 7.2 colour-coded by feature values for selection Gaia DR3 features and scaled to between 0 and 1.

7.3.1.1 PCA Class Projections

The projections with and without the use of Gaia DR3 data (see Figures 7.1 and 7.2) are very similar. SU UMa systems are absent from the far left region when Gaia data is used; this seems to be the only major difference between them. The first two principal components account for only ∼24% of the variance in the dataset (with and without Gaia DR3 features), therefore much information is lost in 2D, possibly leading to a lack of separation of examples into distinct clusters. When the object classes are considered,

some degree of class separation is present, specifically between the Z Cam and SU UMa classes, though significant overlap between classes exists. A transition from long-period to short-period CVs occurs as we shift our focus from the bottom left to the top right of the projections (nova-likes to WZ Sge and AM CVns) whether Gaia DR3 data is used or not. Signs of evolutionary factors are therefore evident in these global structure-centric projections.

7.3.1.2 PCA Feature Projections: Outbursting Characteristics

The feature projections from the case without and with the use of Gaia DR3 data are shown in Figures 7.3 and 7.4, respectively. One may use them as a guide to compare the properties of examples in different regions of each projection.

Amplitude-related features as well as simple variability measures, tend to increase in values from left to right, though with a slight downward inclination. These include *Amplitude*, and *Std* (Figure 7.3). This concurs with the class projections in which nova-like systems (no dwarf nova outbursts) are located farthest to the left and SU UMa systems (with their superoutbursts) reside farthest to the right; U Gems reside somewhat in between with their semi-regular normal dwarf nova outbursts.

Amplitude range-specific features display a class-specific value distribution (Figure 7.3). Outburst amplitudes for Z Cams tend to be smaller than those for U Gems, and smaller still compared to SU UMa systems (Otulakowska-Hypka et al., 2016). This pattern is reflected in the features *npeaks_pt5to1*, *npeaks_1to2*, *npeaks_2to5*, and *npeaks_above5*, which represent the number of peaks in the light curve within specific amplitude ranges. As the amplitudes defining each range increase, the highest values for these features progressively move from the bottom left (for the smallest amplitudes, *npeaks_pt5to1*) to the bottom (for the mid-range amplitudes, *npeaks_1to2* and *npeaks_2to5*), and then to the lower right (for the largest amplitudes, *npeaks_above5*). These trends align with the amplitude characteristics of the different dwarf nova subtypes and their corresponding locations in the projections. Furthermore, the values for the feature that defines the maximum amplitude of peaks between 2 and 5 magnitudes (*amp_2to5_g*) increase diagonally from top left to bottom right, corresponding to the transition from nova-likes to Z Cams and then to SU UMa systems.

Figure 7.3 shows that period (and frequency)-based features, derived using a Lomb-Scargle periodogram, exhibit a trend along the diagonal from bottom left to top right. For instance, the amplitude of the first harmonic of the strongest Lomb-Scargle frequency, *Freq1_harmonics_amplitude_1*, reaches its highest value in the top right of the corresponding subplot. This pattern holds for all harmonics across the various frequencies. This region corresponds to examples where only a single outburst or eruption has been observed in the light curve, typically with an observational timespan covering only that event, leading to fewer data points (*n_obs*). Such light curves tend to produce high-amplitude frequencies due to their simplicity. The main CV types found in this region include WZ Sge, novae, and several SU UMa systems. This is the case in each of the class projections (Figures 7.1 and 7.2). As we shift from the upper right towards the lower left of projections, strong periodic signals are less common, such that more frequently outbursting and lower outburst amplitude sources begin to dominate. Therefore, we shift from SU UMa to U Gem, then Z Cam, until stable non-outbursting systems are predominant. This concurs with the approximate trend seen in the *PeriodLS* feature — the period increasing from bottom left to top right.

The synchronous variability of the g and r band light curves, captured by the *StetsonL* variability index, tends to be higher in regions where semi-regular outbursts occur and where the period measurement is larger (longer), these will be systems with well-sampled outbursts in both bands — SU UMa systems with their superoutbursts are representative of such qualities.

The range of cumulative sum features *Rcs* and *Psi_CS* display a clear trend of increasing values from bottom to top (Figure 7.3). To induce low values of this feature, the light curves display symmetry in points above and below the mean magnitude, such as is seen in frequently outbursting systems with minimal time spent in quiescence (e.g., Z Cams). While high values tend to be exhibited by light curves lacking this symmetry (e.g., dwarf novae with low *n_obs* due to single sampled outburst and VY Scl systems). *LinearTrend*, seems effective in picking out those light curves where only a single outburst/eruption is captured such that the light curve will show a positive magnitude change trend when applying a linear fit, these typically correspond to nova and WZ Sge light curves. *MaxSlope*, the maximum gradient between any two consecutive points, shows an unclear trend, though the highest values tend to align with the rapidly outbursting Z Cams systems as one may expect.

7.3.1.3 PCA Feature Projections: Colour and brightness

ZTF colour related features (*clr_mean*, *clr_bright* and *clr_faint*) do not show any clear trend in Figure 7.3. Neither do the Gaia colours, *bp_rp*, *bp_g*, and *g_rp* shown in Figure 7.4. All that can be said is that ZTF colour values show a faint trend decreasing in the direction from the lower left to the upper right. This aligns with the upper right belonging to some of the shortest periods and bluest systems (see Figures 7.1 and 7.2). Gaia absolute magnitudes, apparent magnitudes and fluxes (e.g., *absmag_g*, *phot_g_mean_mag* and *phot_bp_mean_flux*) show a clear linear trend corresponding to a transition from the intrinsically or apparently brightest systems to those that are intrinsically or apparently faintest (e.g., nova-likes to SU UMa and AM CVns).

7.3.1.4 Magnetic CVs

The analysis so far has yet to touch upon the population of magnetic CVs, namely the intermediate polars (IPs) and polars. The IPs reside farther left than the polars, in both the with and without Gaia DR3 representations, though much scatter is present for both classes. Little more can be said of the location of these objects but that the lack of locality may just point to the diversity of light curve profiles. Such profiles include outbursts in several IPs, long-term fluctuations in brightness due to mass-transfer rate changes (more prominent in polars than from IPs), and eclipses from periodic occultations of different parts of the accretion geometry. Diversity in light curve profiles may also play a factor in the lack of locality of AM CVs.

7.3.1.5 Factors Impacting Projections

Many of the feature projections show a clear and linear trend, these tend to provide the greatest clarity in interpreting the class projections. However, several features show significant scatter, providing less clarity for interpretation. One significant factor would be the wide range of orbital inclinations. This impacts multiple measurable properties, including the depth of eclipses, and the apparent and absolute brightness, which is also a function of wavelength, colour measurements, and outburst amplitudes. Another factor could simply be the need for non-linear dimensionality reduction analysis to assess the relevance of such features. Non-linear algorithms are now explored.

7.3.2 t-SNE

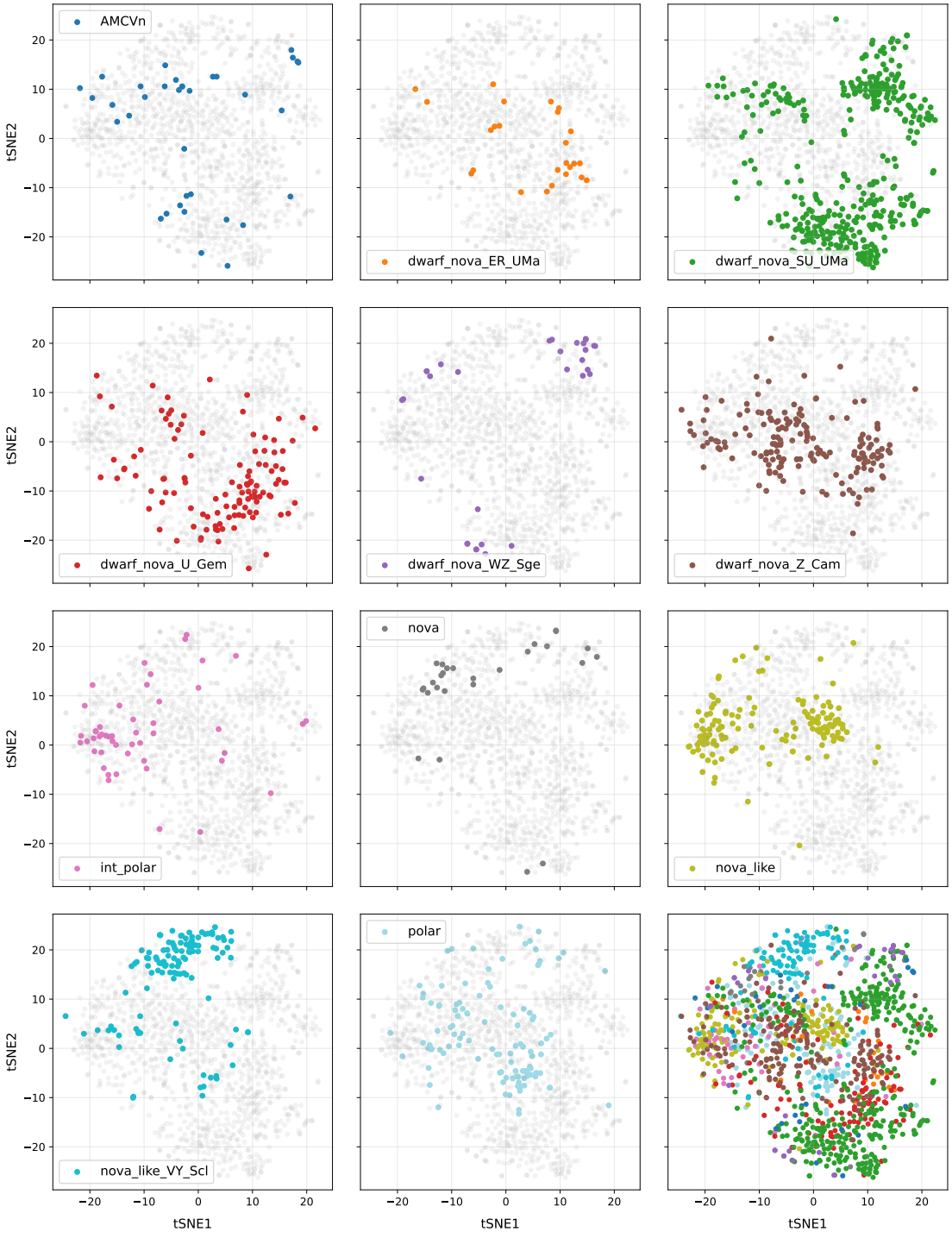


FIGURE 7.5: t-SNE 2D projection of dataset where a minimum points threshold of 20 in either the g or r band was set and no external (DR3) data was utilised. They are colour-coded by class, and presented in a one-versus-rest manner apart from the plot on the bottom right, which combines the preceding plots. The hyperparameters for the model were set as follows: perplexity=20, learning_rate=10, n_iter=1e6, early_stopping=1000, early_exageration=12

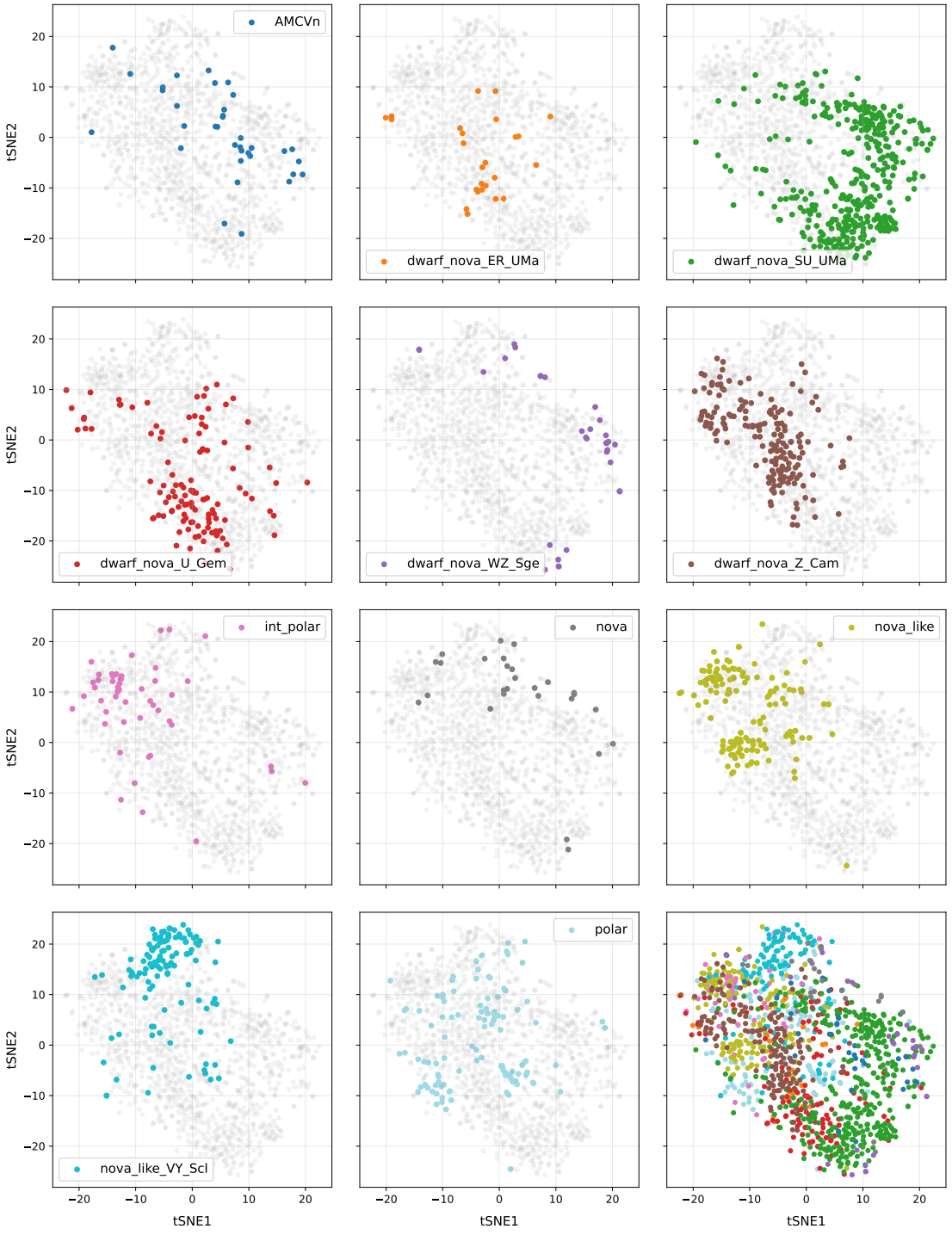


FIGURE 7.6: Same as figure 7.5 but with external (DR3) data. Hyperparameters are as follows: perplexity=20, learning_rate=10, n_iter=1e6, early_stopping=1000, early_exageration=12

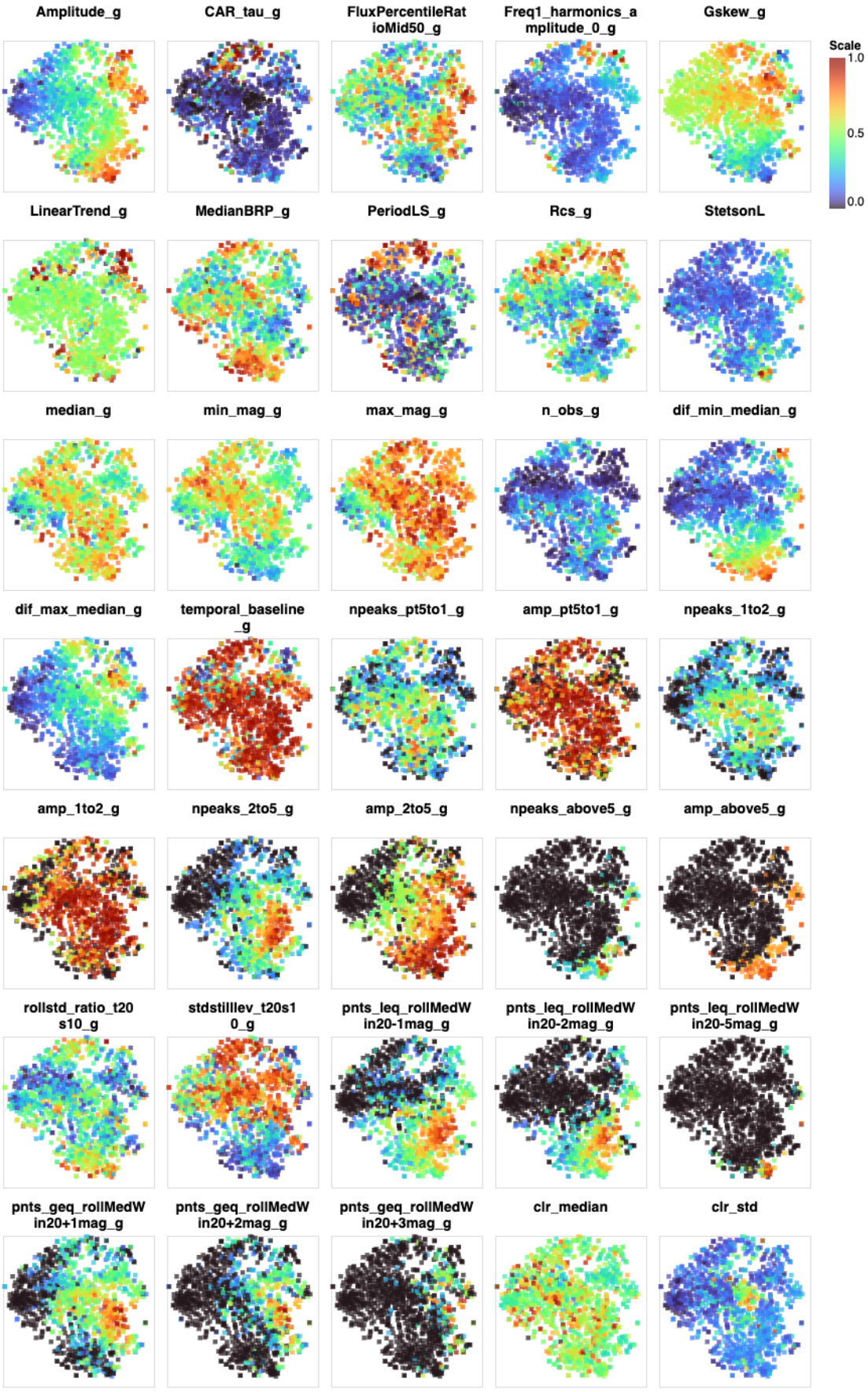


FIGURE 7.7: t-SNE projections without inclusion of Gaia DR3 data (see Figure 7.5) colour coded by feature values for selected features and scales to between 0 and 1.

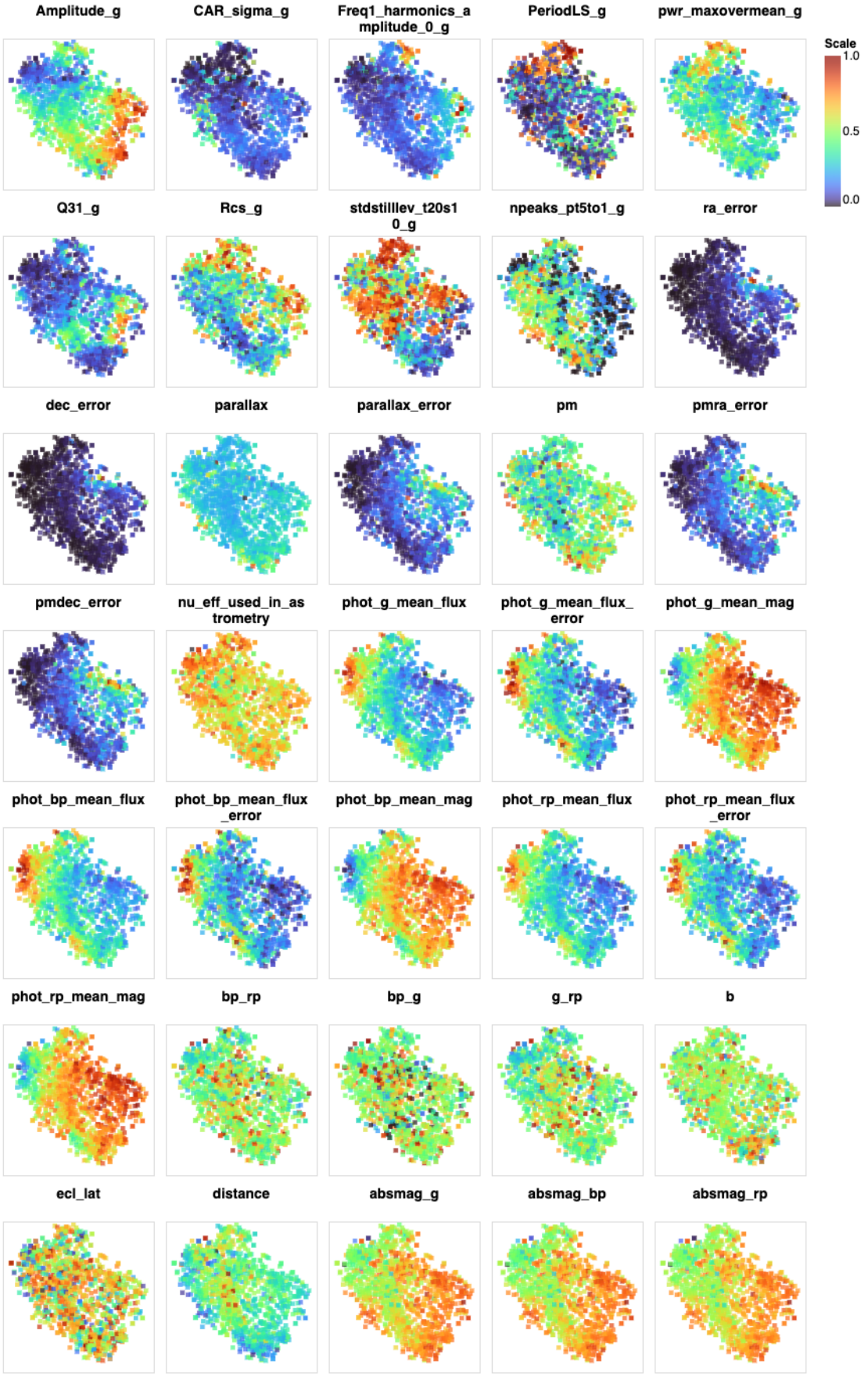


FIGURE 7.8: t-SNE projections with inclusion of Gaia DR3 data (see Figure 7.6 colour-coded by feature values for selection Gaia DR3 features and scaled to between 0 and 1.

7.3.2.1 t-SNE Implementation

To optimise the projections with t-SNE, the perplexity was adjusted in increments of 5 from 5 to 100, with subsequent adjustments of the learning rate and early exaggeration parameters for fine-tuning. The choice of optimal parameters was based upon the degree of cluster separation, and the repeatability of the separation of examples belonging to specific classes under different random initialisation states. The projections from the use of t-SNE, both without and with Gaia DR3 data (Figures 7.5 and 7.6), display a separation of examples into clusters of higher example densities, unlike PCA. The cluster separations are not as clear as one would hope, with a fair amount of blending. However, when considering the class labels, one can see clear and almost distinct locations for several of the classes. During the optimisation process, it became evident that consistent patterns concerning the separation of classes were always present regardless of hyperparameter values, as I will now describe.

7.3.2.2 Nova-likes

Gaia DR3 data or not, there always exist two distinct regions heavily populated by nova-like systems (olive green in Figures 7.5 and 7.6). One region belongs to examples whose light curves display clear eclipses with depths typically a magnitude or greater. Other classes also reside within or very close to this region, all of which exhibit signs of eclipsing behaviour.

Referring to the feature projections of Figure 7.7, one can see that high values of the standard deviation of colour (*clr_std*) are associated with this eclipsing region. Eclipses impact the g and r bands differently based on the components being obscured. Typically, eclipses of the accretion disk lead to redder colours, while eclipses of the secondary star result in bluer colours. The interplay between these effects in CVs produces complex colour variability, which can be characterised by *clr_std*. Low values of the percentile range ratio *FluxPercentileRatioMid50* are also associated with this region. This ratio provides insight into the distribution of flux values, specifically how concentrated or spread out the majority of the flux values are within the light curve. A low ratio indicates that the flux variation within the central 50th percentile range is relatively small compared to the total flux variation, i.e., that most of the light curve's variation

is concentrated at the extremes. In this case, caused by the sharp flux drops during eclipses and relatively stable flux levels outside of them.

The other nova-like region is always populated by several Z Cams and a concentrated area of intermediate polars, where the Z Cams and intermediate polars usually display some separation. Here the light curve variability is restricted to within 1 magnitude or less. Consequently, low values of *Amplitude* are a marker for this region. For this region, the Z Cams possess low amplitude high-frequency outbursts, while where intermediate polars are most prominent, a slightly more stochastic variability takes place.

7.3.2.3 VY Scl systems

VY Scl systems (cyan in Figures 7.5 and 7.6) are well separated from other classes regardless of the use of Gaia DR3 data. Referring to feature projections without Gaia DR3 data (Figure 7.7), the region is associated with the longest Lomb Scargle period (*PeriodLS*) presumably due to long intervals between low state excursions; and a high value for the standstill level, *stdstilllev_t20s10*, a consequence of the long featureless high states. Interestingly, one may divide this region into two based upon *Freq1_harmonic_amplitude_0*, where high values correspond to an almost constant light curve with only the beginning or the end of a single low state excursion present, while low values are associated with multiple or complete low state excursions.

7.3.2.4 Polars

For the projections without Gaia DR3 data, a distinct concentration of polars is consistently observed, separate from other classes (just below the centre in Figure 7.5). While some contamination from VY Scl systems exists, the variability in their light curves closely resembles that expected from polars. In the feature projections (Figure 7.7), this region is associated with high values of *PeriodLS*, potentially reflecting long-term high and low state fluctuations. Additionally, it exhibits a pocket of relatively low *MedianBRP* values, which represent the fraction of data points within amplitude/10 of the median magnitude, indicative of high variability.

When Gaia DR3 data is incorporated (Figure 7.8), two relatively distinct clusters of polars emerge, further separating them from other classes. These clusters differ primarily

in the amount of variance present in the light curves. The cluster furthest to the left in Figure 7.6 is characterised by significantly higher variance (short timescale variability), whereas the other cluster, located slightly lower and to the right of the centre, exhibits less variance. This distinction is best captured by the *CAR_sigma* feature, as shown in Figure 7.8.

7.3.2.5 Z Cams

Three clusters of Z Cam systems typically emerge, one smaller than the other two, regardless of whether Gaia DR3 data is included. These clusters are not well-defined, displaying significant scatter and blending between them. The two larger clusters are generally positioned on either side of the eclipsing nova-like region mentioned earlier, with a portion of one of these clusters overlapping with a significant number of U Gem systems.

In the case without Gaia DR3 data (Figure 7.5), the two larger clusters are located on either side of the centre, while the smaller cluster is situated furthest to the left. When Gaia DR3 data is used (Figure 7.6), the two larger clusters tend to blend more, lying on either side of a population of nova-likes, with the smaller cluster shifting toward the top left. Among the larger clusters, one is characterized by higher amplitude outbursts compared to the other. This distinction is best captured by features such as the number of peaks with amplitudes between 2 and 5 magnitudes (*npeaks_2to5*) and the number of data points falling below (or above) the median of a rolling median window with a 1-magnitude threshold (*pnts_leq_rollMedWin20-1mag* or *pnts_geq_rollMedWin20+1mag*) (Figure 7.7).

The smaller Z Cam cluster, as noted earlier, is mixed with low-amplitude nova-like systems and exhibits low-amplitude outbursts or variability. However, the clear separation of sources based on the presence of standstills remains elusive. Where standstills are present, they appear to be more pronounced in the lower-amplitude large Z Cam cluster.

7.3.2.6 SU UMa systems

SU UMa systems occupy several regions in the projections, regardless of whether Gaia DR3 data is included (Figures 7.7 and 7.8, respectively). One of these regions overlaps

with a significant number of U Gem systems, which is unsurprising given the challenge our features face in distinguishing these two classes. This difficulty arises primarily from the similarity between superoutbursts and normal outbursts, which can often be hard to differentiate.

The distinct SU UMa regions, whether or not Gaia DR3 data is used, appear to be separated based on factors such as the level of quiescent sampling, outburst amplitude, and outburst recurrence period. For example, in the projection without Gaia DR3 data (Figure 7.5), the concentrated SU UMa region around tSNE1 = 10 and tSNE2 = 10 is populated by systems with high outburst amplitudes and well-sampled quiescence. Consequently, the *Amplitude* feature shows high values in this region, while the standstill level (*stdstilllev_t20s10*) is very low.

Moving leftward in this projection, outburst recurrence periods increase until reaching a population dominated by WZ Sge systems. These systems typically exhibit only a single observed outburst within the ZTF survey's observational timespan but maintain good quiescent sampling. In the northernmost SU UMa region, systems are found where no quiescent observations have been recorded; only outbursts have been observed. Here, the *Amplitude* feature remains high due to the lack of quiescent sampling, and the *stdstilllev_t20s10* feature also shows high values in this region. Additionally, in the upper-right portion of the projection without Gaia DR3 data, another group of WZ Sge systems appears, again lacking quiescent sampling.

When Gaia DR3 data is included (Figure 7.6), the SU UMa-dominated regions exhibit similar attribute-based separations, despite differences in their specific locations within the projection. High outburst amplitude, recurrence period, and the extent of quiescent sampling continue to define the substructure of these regions.

7.3.2.7 ER UMa and U Gem

ER UMa systems are short-period, high mass-transfer rate dwarf novae. As a result, their supercycles (the time between successive superoutbursts) are exceptionally short (less than 50 days), as are their normal outburst cycles. These rapidly outbursting systems often overlap in location with Z Cam systems that also exhibit frequent outbursts.

Consequently, neither the projection without Gaia DR3 data nor the one with it provides a distinct or consistent location for ER UMa systems.

U Gem systems, regardless of random initialisation of the algorithm or the inclusion of Gaia DR3 data, are predominantly found in a large, loosely defined cluster that links Z Cam and SU UMa systems. Smaller numbers of U Gem systems also appear in several other diffuse regions that are shared with other dwarf nova varieties. This distribution suggests that the light curves and, where used, Gaia DR3 data do not provide sufficient discriminative power to fully separate U Gems from Z Cam and SU UMa systems. This is unsurprising, as inspecting the dwarf nova light curves in our dataset reveals significant overlap in outburst recurrence periods and amplitudes between these classes.

Nevertheless, the majority of U Gem systems in both Figure 7.5 and Figure 7.6 are associated with sources exhibiting the highest number of outbursts (peaks) with amplitudes between 2 and 5 magnitudes. This trend is evident in the *npeaks_2to5* feature projection (Figure 7.7). Such behaviour aligns with a defining characteristic of U Gems: their outburst amplitudes typically fall within this range (Otulakowska-Hypka et al., 2016).

7.3.2.8 Novae

Novae predominantly occupy regions characterised by high values of the *LinearTrend* feature, as illustrated in Figures 7.5 and 7.7. This feature likely captures the monotonic decline brightness typical of post-eruption light curves. Additionally, novae are associated with low values of *n_obs* (the number of data points in the light curve), reflecting their limited observational window, which is often restricted to their eruption phase. In Figure 7.5, novae are dispersed within the top third of the projection, a spread that appears to correspond to variations in light curve amplitude.

When Gaia DR3 data is incorporated (Figures 7.6 and 7.8), novae cluster in regions associated with high errors in positional coordinates, parallax, and proper motion (*pmra_error* and *pmdec_error*). These high errors can be attributed to several factors: the absence of pre-outburst observations, their typical placement at significantly greater distances than other CV subtypes, and their preferential location at low Galactic latitudes. At these latitudes, interstellar extinction is more pronounced, further complicating accurate astrometric measurements.

7.3.2.9 AM CVns and Other

AM CVn systems do not exhibit strong localisation in either projection, whether Gaia DR3 data is included or not. Several factors may contribute to this lack of clustering: the limited number of known examples, the wide range of variability patterns arising from their evolutionary stages, their intrinsic faintness, and variations in orbital inclination. However, in the projection with Gaia DR3 data (Figure 7.6), there is a slight indication of localisation near coordinates (tSNE1=10, tSNE2=-5). One might expect this region to correspond to systems with high proper motion (*pm*) and/or low *bp_rp* colour values (indicating bluer systems), but no distinctive trends in these features are evident in this area (see Figure 7.8).

Both projections reveal regions characterised by a mixture of different classes. In the projection without Gaia DR3 data (Figure 7.5), this region lies northwest of the centre, while in the projection with Gaia DR3 data (Figure 7.6), it is located north of the centre. These regions are associated with light curves that have relatively low amplitudes and sparse sampling (*n_obs*), reflecting cases where the data available makes it particularly challenging to assign a definitive classification.

7.3.3 UMAP

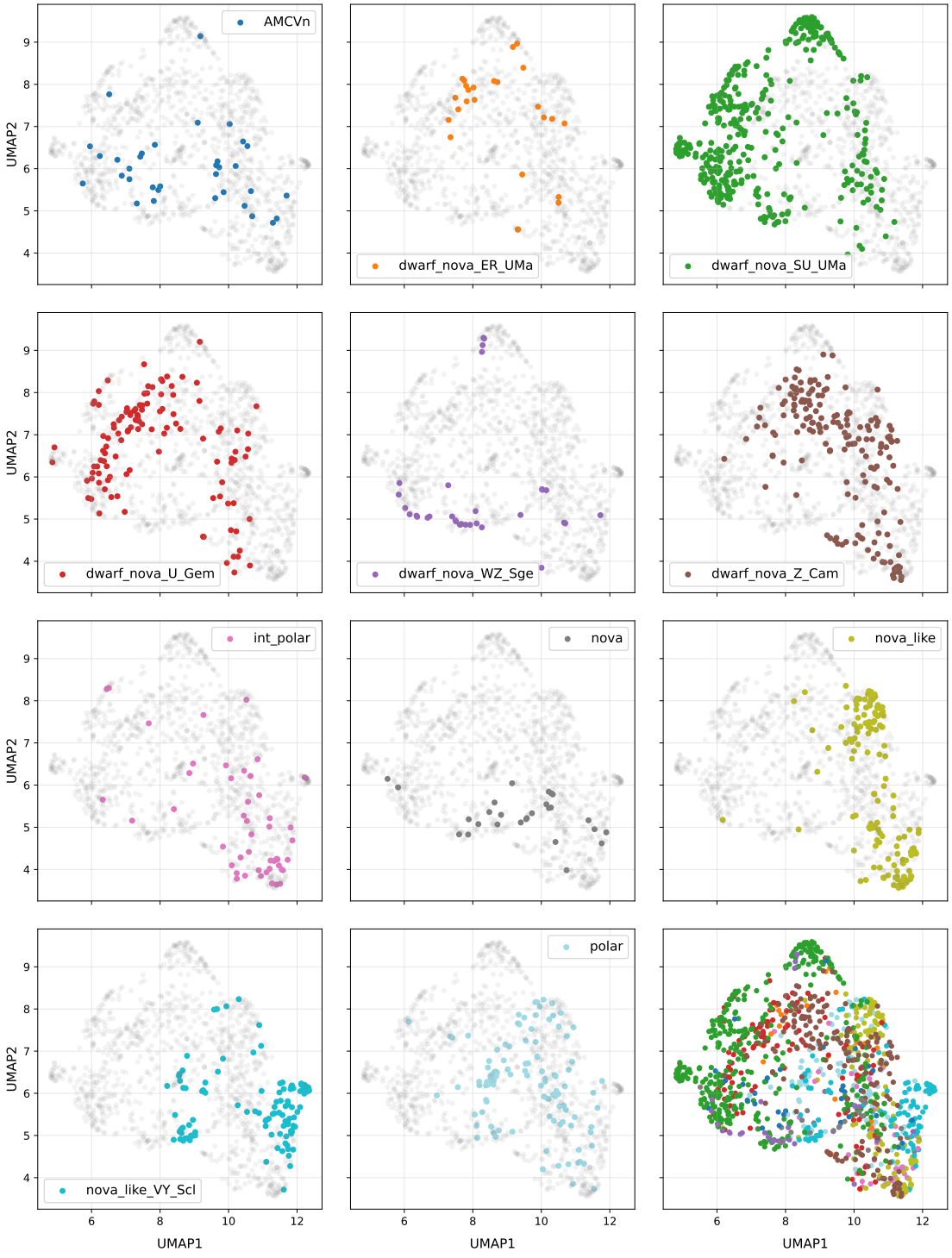


FIGURE 7.9: UMAP 2D projection of dataset where a minimum points threshold of 20 in either the g or r band was set and no external (DR3) data was utilised. They are colour-coded by class, and presented in a one-versus-rest manner apart from the plot on the bottom right, which combines the preceding plots.

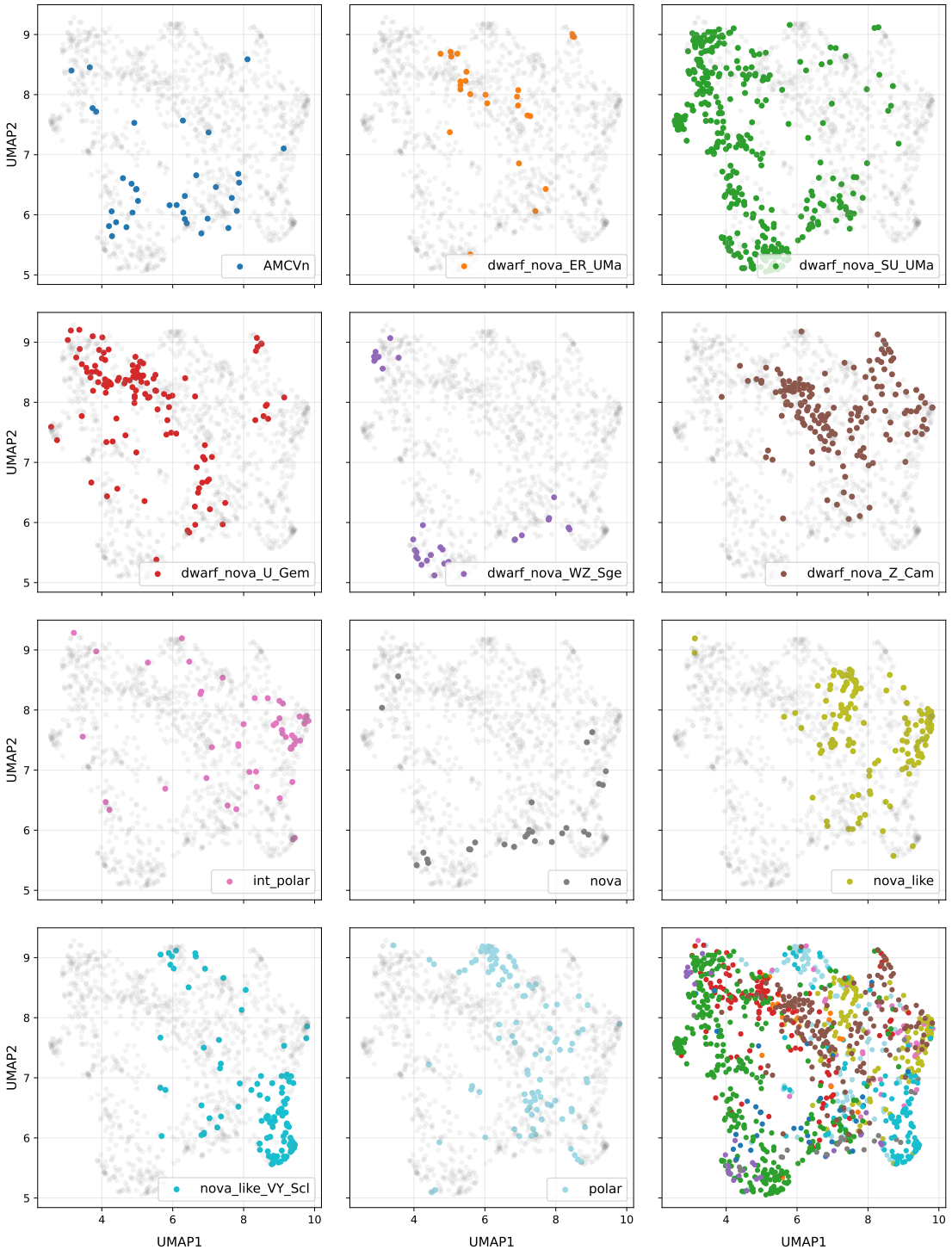


FIGURE 7.10: Same as Figure 7.9 but with external (DR3) data.

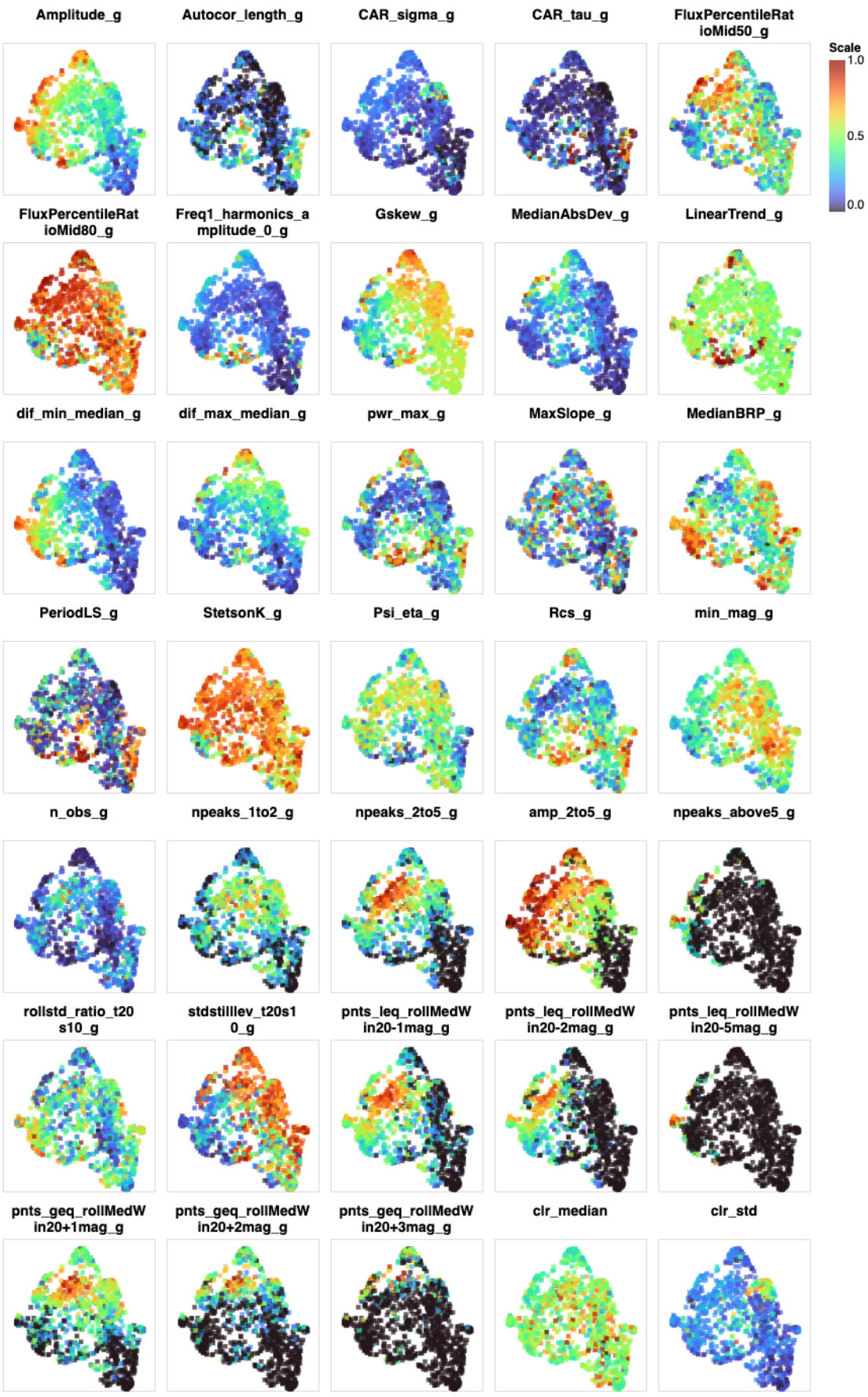


FIGURE 7.11: Feature projections for UMAP without the use of Gaia DR3 data

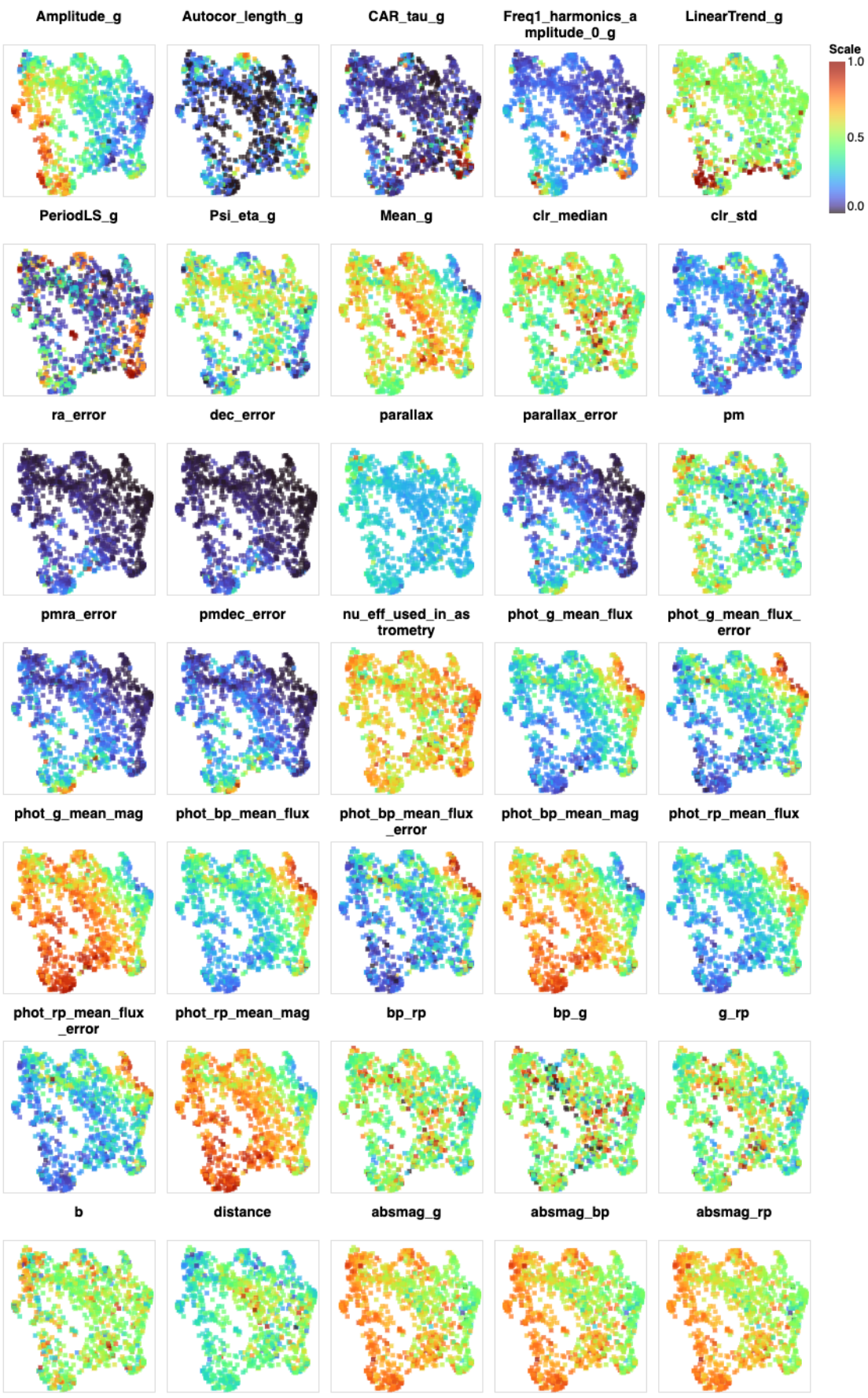


FIGURE 7.12: Feature projections for UMAP with the use of Gaia DR3 data

7.3.3.1 UMAP implementation

To optimise the UMAP projections, the objective was to create representations with as distinct clusters as possible. However, as anticipated, some blending between classes is present in the projections, both with and without Gaia DR3 data (Figures 7.9 and 7.10, respectively). The *n_neighbours* hyperparameter had the most significant influence on the projections. For each value of this hyperparameter, projections were also generated using various values of the *learning_rate* and *min_dist*. As *n_neighbours* approached 100, the local structure diminished, while values in the single digits resulted in many small clusters (typically 2–3 examples). Increasing *n_neighbours* allowed for a better balance between local and global structure, with further fine-tuning achieved using the *learning_rate* and *min_dist* parameters. For the projection without Gaia DR3 data, an *n_neighbors* value of 25 appeared to be optimal, whereas for the projection with Gaia DR3 data, a value of 30 produced a satisfactory result. Learning rates of 0.01 were used for both sets of projections, while a *min_dist* of 0.05 was used for the non-Gaia DR3 case and 0.01 for the projection with Gaia DR3 data. To test the repeatability of the optimisation process, projections were generated with different random initialisations using the same hyperparameters. Regardless of the random seed, the same clustering patterns consistently emerged, reinforcing the reliability of the optimisation procedure.

7.3.3.2 Comparisons with PCA and t-SNE

Projections generated with UMAP show dense clusters with a clear separation between them, as seen in Figures 7.9 and 7.10. The separation in UMAP appears more pronounced than in t-SNE (and more than in PCA), although this is highly sensitive to the choice of hyperparameters for both t-SNE and UMAP. Additionally, UMAP tends to exhibit more filamentary connections between clusters, as opposed to the scattered blending observed with t-SNE. The differences between UMAP projections with and without Gaia DR3 data are minimal, though these may be influenced by the random initialisation state and specific hyperparameter choices.

Upon adding class labels, the UMAP projections reveal the same class segregation patterns observed in the t-SNE projections. Notably, two nova-like regions emerge: one with eclipsing systems and another with less variability, which contains a significant group of

intermediate polars along with several Z Cams. Several well-separated SU UMa clusters appear distinct from other classes. ER UMa systems align with Z Cams, and Z Cams surround the eclipsing nova-like systems in two larger clusters. WZ Sge systems occupy two or three smaller regions located at the outer edges of the SU UMa clusters. A large, well-separated cluster of VY Scl systems is also evident. U Gems overlap with some SU UMa and Z Cams, creating a link between these two classes. There is a lack of localisation for AM CVn systems. One polar cluster is clearly separated from other classes when no Gaia DR3 data is used, while two distinct polar clusters are present when Gaia DR3 data is incorporated. These characteristics are consistent across both projections with and without Gaia DR3 data. The primary difference between the t-SNE and UMAP projections lies in the relative locations of these clusters in the respective projections. A few other notable differences also appear.

The two polar clusters in the Gaia DR3 projection differ slightly between t-SNE and UMAP (Figures 7.6 and 7.10). In t-SNE, they are well-separated, with minimal contamination, and distinguished by short-timescale variability. In UMAP, one cluster is more distinct, while the other overlaps with nearby classes.

In the UMAP projection without Gaia DR3 data (Figure 7.9), VY Scl systems split into two clusters: a larger one at the bottom right and a smaller one to the left. The feature *Freq1_harmonics_amplitude_0* shows lower values for the systems in the bottom-right cluster, indicating clearer periodicity in the left cluster. Light curve inspection confirms that the left cluster contains systems better described by a periodic function.

With Gaia DR3 data (Figure 7.10), VY Scl systems are mostly concentrated in the bottom-right region, with some subcluster separation. The *Freq1_harmonics_amplitude_0* feature captures this separation: higher values appear in the southernmost examples, where a periodic function fits well the associated light curves — often marking the start or end of a low-state excursion. Moving north, the full profile of low-state excursions is present in the light curves. The most northerly examples of the VY Scl systems in the bottom-right region of the projection show shallow low-state excursions.

7.3.3.3 Feature Projections

When comparing the feature projections of UMAP with and without Gaia DR3 data (Figures 7.11 and 7.12) to the corresponding class projections (Figures 7.9 and 7.10), the associations between feature values and class locations observed in t-SNE are also evident in UMAP. However, some additional observations can be made.

For projections with Gaia DR3 data, the highest values of Galactic latitude (b) are found in regions where short-period systems, such as SU UMa and WZ Sge, are located. This region also corresponds to some of the closest systems. The light curves of these systems are well-sampled during quiescence. One explanation for this is that, during quiescence SU UMa and WZ Sge are some of the intrinsically faintest objects in the optical and would be required to be close by to be observable, this is more so the case here due to a well-sampled quiescence. Furthermore, for nearby sources, the range of latitudes that are 'in the Galactic plane' is higher than for more distant sources.

Another observation is the Gaia colours. A streak of redder colours (higher values) in the central regions of the feature projections for *bp_rp*, *bp_g*, and *g_rp* (Figure 7.12) is associated with longer-period systems, such as Z Cams and nova-likes. Shorter-period systems are associated with bluer colours (lower values). While the projections do not fully clarify this, a similar pattern is discernible.

7.3.4 Generative Topographic Mapping

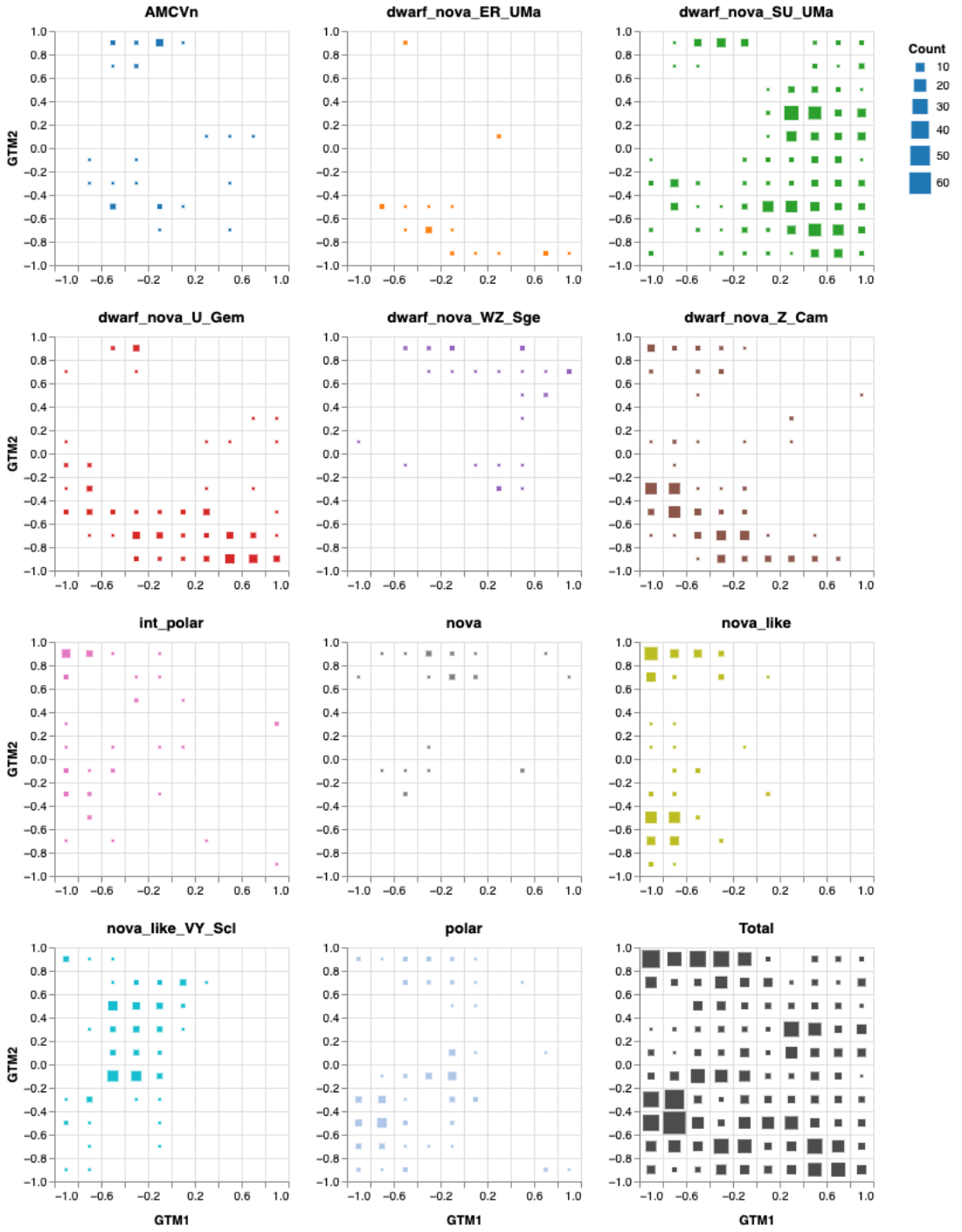


FIGURE 7.13: GTM 2D projection of dataset where a minimum points threshold of 20 in either the g or r band was set and no external (DR3) data was utilised. They are colour-coded by class, and presented in a one-versus-rest manner apart from the plot on the bottom right, which combines the preceding plots.

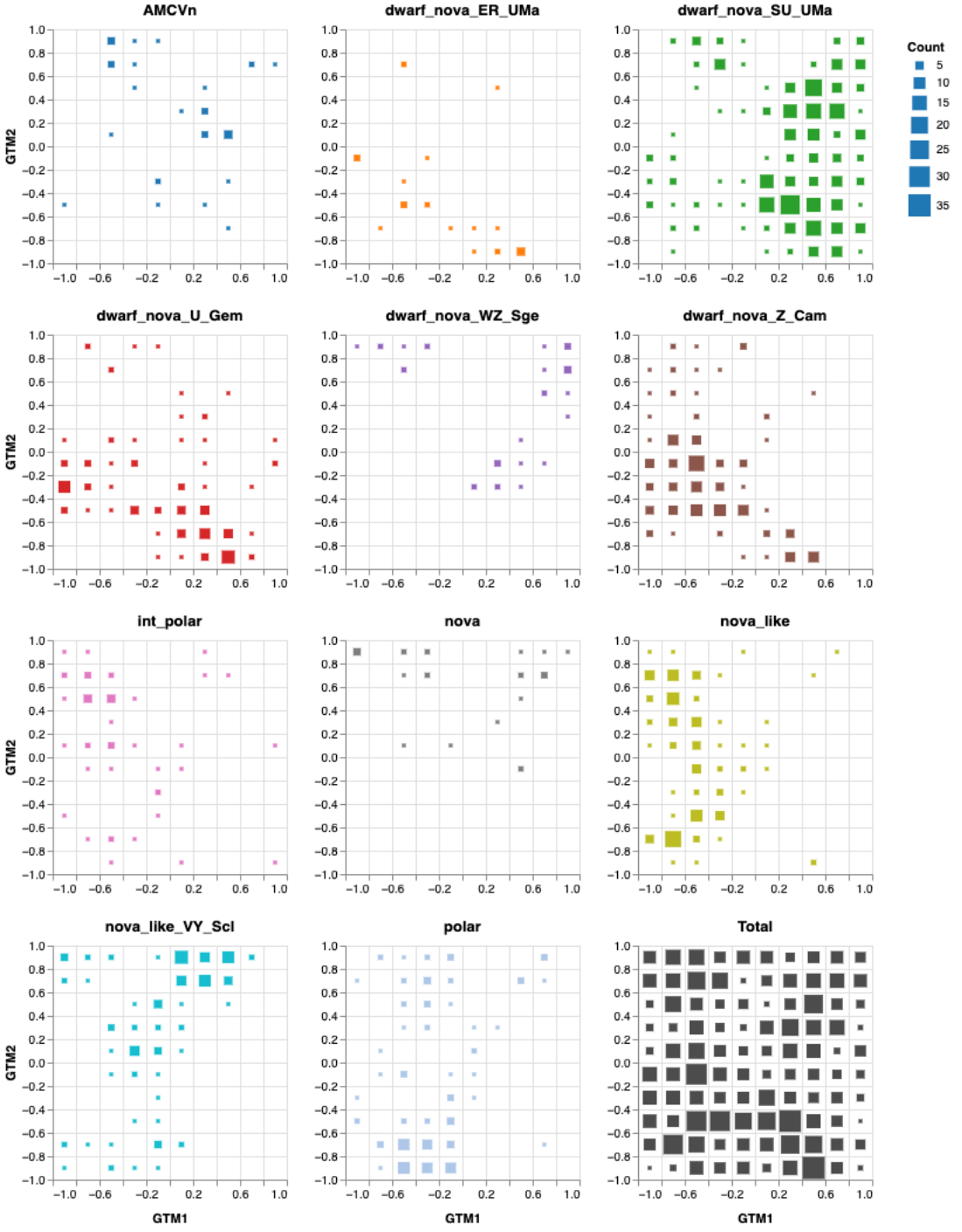


FIGURE 7.14: Same as figure 7.13 but with external (DR3) data.

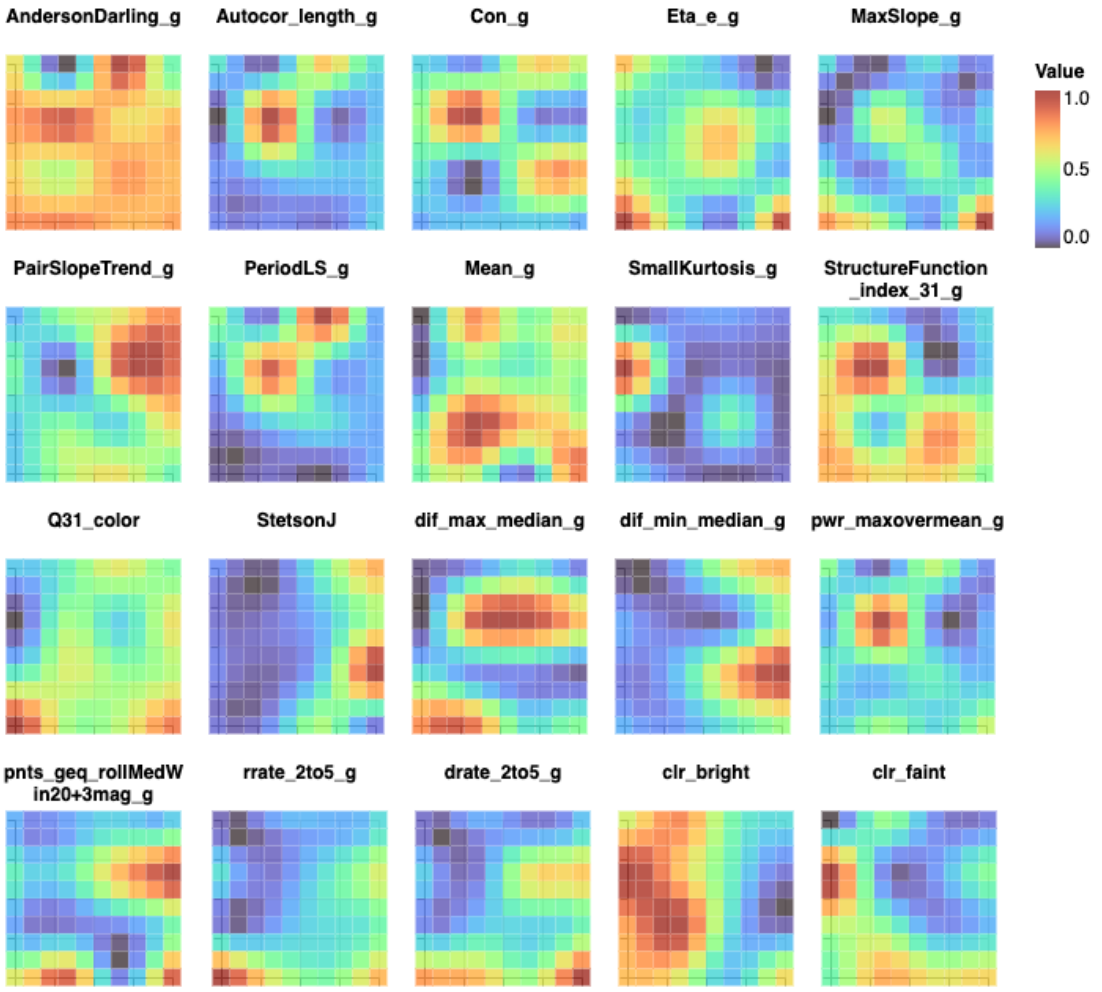


FIGURE 7.15: Reference maps for the GTM projection without the use of Gaia DR3 data for several features.

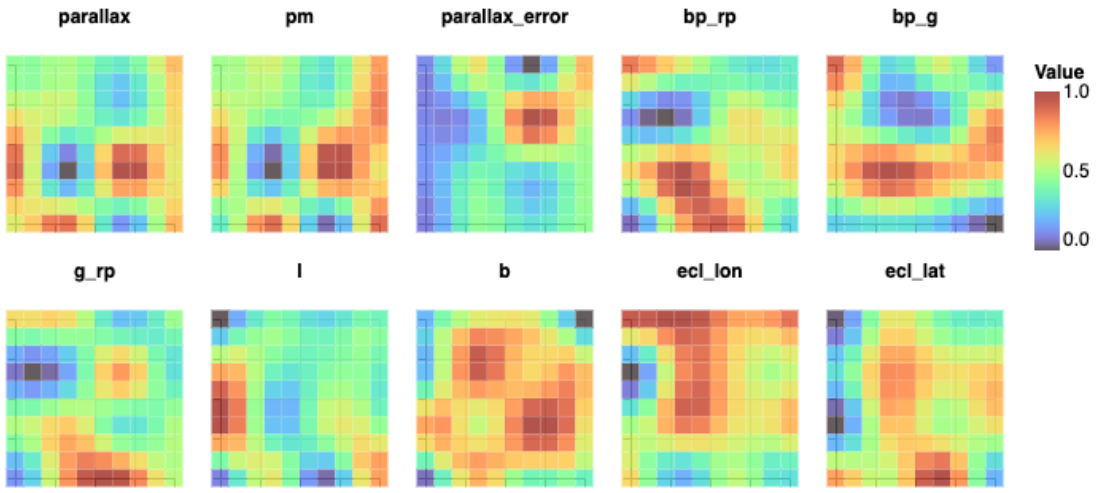


FIGURE 7.16: Reference maps for the GTM projections with the use of Gaia DR3 data.

7.3.4.1 GTM Implementation

The optimisation of GTM projections primarily involved adjusting the square root ($\sqrt{}$) of the number of GTM nodes (k) for the latent space representation. Hyperparameters related to the square root of the number of radial basis function (RBF) centres (m), the RBF width factor (s), and the regularisation coefficient ($regul$) had minimal to no discernible impact on the resulting latent space projections.

Overfitting may occur when the number of nodes approaches the number of examples, as many nodes remain unoccupied, providing no meaningful contribution to the feature space representation. Conversely, underfitting can arise with too few nodes, leading to an inability to capture the underlying patterns in the dataset.

A configuration of 100 nodes ($k = 10$) struck a balance, ensuring that most nodes were assigned examples while minimising the number of empty nodes. This configuration proved effective for both cases — with and without the inclusion of Gaia DR3 data.

7.3.4.2 Class Projections

Figures 7.13 and 7.14 are the projections without and with the use of Gaia DR3 data. These are the mode projections, such that examples are assigned to the most responsible node. The figures appear as histograms, providing an indication of the number of examples assigned to each node, also separated by class. The distribution of the different CV classes follows a similar theme to the distributions for t-SNE and UMAP, e.g., two nova-like regions, one or two regions of polars; a well-defined region of VY Scl; U Gems providing a link between SU UMa and Z Cams; etc. This helps to confirm that the projections with t-SNE and UMAP are not just a consequence of a particular set of hyperparameters or particular random state initialisation, but are a sound representation of the high dimensional distribution.

7.3.4.3 The Advantage of GTM and Reference Maps

GTM offers a significant advantage over PCA, t-SNE, and UMAP by enabling the creation of reference maps, which provide a clearer and more interpretable view of feature

space. Unlike the projections from t-SNE and UMAP, which are sensitive to hyper-parameter choices and can obscure relationships due to class overlap in dense regions, GTM's probabilistic framework offers a structured perspective on the data distribution.

Each GTM node represents a Gaussian centred on a cluster of data points (CVs) in the high-dimensional space, thereby modelling the probability distribution of the data. The reference maps visualise this distribution by explicitly showing the positions of these Gaussian centres in feature space, facilitating the interpretation of the properties of each cluster of CVs and the relationships between them. Additionally, by providing a probabilistic view of the data, the reference maps allow for a more nuanced understanding of the dataset's diversity.

While GTM reference maps do not directly quantify the influence of individual features, discernible structures within these maps can suggest qualitative relationships between features and data distribution. Features that appear unstructured in t-SNE or UMAP projections can show clear structure in GTM reference maps (Figures 7.15 and 7.16), therefore, reference maps can serve to highlight their relevance in defining CV properties.

7.3.4.4 Reference Maps

High values of *AndersonDarling* indicate that the magnitude distribution in a light curve significantly deviates from normality. In the reference map (Figure 7.15), the highest values correspond to light curves capturing only the decline of a single outburst or eruption, resulting in skewed distributions. VY Scl systems also exhibit skewed magnitude distributions and are well-represented by this feature. In contrast, low values are associated with nova-likes or sources with nearly constant or minimally structured light curves.

The *Autocor_length* feature measures how quickly autocorrelation decays as the time lag increases. VY Scl systems exhibit the longest *Autocor_length* values, reflecting extended correlation over time due to their sustained brightness states. In contrast, shorter values correspond to systems with short-term variability, such as frequently outbursting dwarf novae.

The *Con* feature characterises the number of three consecutive points brighter or fainter than 2σ of the light curve mean, normalised by the number of data points. Long and well-sampled phases of constant brightness, interspersed with significant and well-sampled deviations from this induce high values. Light curves of nodes where this value is highest exhibit these characteristics — VY Scl, and SU Uma stars with quiescent sampling. The lowest values tend to be dwarf nova without quiescent sampling and Z Cams with poorly sampled or short standstills.

The *Eta_e* feature quantifies the independence of successive data points in a light curve. High values correspond to systems with rapid, short-timescale variability, such as frequently outbursting dwarf novae and strongly eclipsing systems. Lower values are linked to smoothly varying light curves, like those with only a single sampled outburst or eruption. However, *Eta_e* also accounts for the time differences between data points, so sparse sampling reduces the feature's value. For example, nodes in the middle-bottom of the reference map, where values are low (Figure 7.15), correspond to sparsely sampled Z Cam systems. In contrast, nodes on the bottom right, where values are high, represent highly variable and well-sampled dwarf novae.

The *MaxSlope* feature represents the steepest slope between consecutive points in a light curve. As expected, the lowest values are observed for low-amplitude systems, while the highest values occur in rapidly outbursting but well-sampled dwarf novae.

The *PairSlopeTrend* feature measures the trend (increasing or decreasing) in the last 30 magnitude measurements (or fewer if the light curve contains fewer points). Negative gradients can be observed when VY Scl systems exit low states, while positive gradients occur when only the decline from an outburst or eruption is sampled. This general trend aligns with the light curves of nodes where the feature has extreme values.

7.3.4.5 Features Revealing Structure in GTM Maps

The Lomb-Scargle period (*PeriodLS*) reference map better captures light curve periodicity than the t-SNE and UMAP feature projections. The *Mean* (mean magnitude) reference map also better aligns with the typical brightness of the CVs. While *SmallKurtosis* shows no structure in t-SNE and UMAP projections, it displays clear structure in the reference map; nodes with low and high values distinguish between light curves

with less and more extreme variability, respectively. The structure function measures the variability of a time-series signal as a function of the time separation between data points (or time lag). *StructureFunction_index_31* in particular measures the slope of the structure function over 1- and 3-day time lags, reflecting how variability changes across timescales. The associated light curves show that long-term smooth changes induce the lowest values, while short-timescale variability induces higher values.

The difference between the third and first quartile of the epochal colours (*Q31_color*) is similar to the *clr_std* feature but less sensitive to extreme colour values. The reference map structure highlights strong eclipsers and frequently outbursting dwarf novae with well-sampled outbursts at their highest values. *StetsonJ* quantifies the correlation between magnitudes in the g and r bands, reflecting synchronicity in variability. This feature is most pronounced when the g and r band light curves vary in sync during outbursts, as seen in SU UMa with well-sampled superoutbursts.

The features *dif_max_median*, *dif_min_median*, and *pnts_geq_rollMedWin20+3mag* enhance the amplitude feature by further categorizing high-amplitude light curves based on brightness deviations from the average. The reference maps for rise and decline rates, *rrate_2to5* and *drate_2to5*, effectively distinguish between rapidly changing and slowly varying light curves. The *clr_bright* and *clr_faint* reference maps show clear structural differences, highlighting how varying stages of activity affect colour measurements.

7.3.4.6 Gaia DR3 Reference Maps

When Gaia DR3 data is used, both parallax and proper motion (*pm*) show clear, similar structures in the reference maps. The distribution of examples aligns with the most distant sources, such as nova-likes and Z Cams having the lowest parallax and proper motion, and the closest sources, such as the SU UMa and the WZ Sge subtype, having the highest values — required to be closer to be observable due to their optical faintness relative to nova-like and Z Cams.

Gaia colour features (*bp_rp*, *bp_g*, and *g_rp*) display clearer structure in the reference maps compared to t-SNE and UMAP projections. These features provide insights into the relative contributions of different CV components, such as the white dwarf, accretion disk, and donor.

The level of dust and interstellar extinction, the spatial distribution of stellar populations (such as their age and metallicity), and the coverage of sky regions by surveys all vary depending on the sky location. These factors can influence the observed properties of CVs, such as their apparent brightness and colour. The sky location is defined using Galactic and ecliptic coordinates (l, b, *ecl_lon*, and *ecl_lat*). Given these varying conditions, one might expect location-based influences to result in distinct structure in the feature projections of t-SNE and UMAP, however, such structure is only present for the GTM reference maps. This facilitates further investigation that would not be possible with the t-SNE and UMAP projections.

Finally, the parallax error, *parallax_error*, is structured in such a way that the highest values are associated with the faintest sources; they tend to be systems only observable during heightened activity (high state or outbursts).

7.3.5 Projection of new examples

Of all the ZTF CVs identified by a cross-match with the AAVSO CV list that match our data point threshold criteria, 236 have only a broad CV classification and 2,170 have only a broad dwarf nova classification, with no further granularity. From here on, I refer to these sources as out-of-sample CVs and out-of-sample DNe, respectively. In this section, I show where these out-of-sample sources lie within the learnt UMAP projections generated with and without Gaia DR3 data. Then using the training set examples as a guide I aim to provide a more granular classification for the out-of-sample sources. I perform this for several locations and produce a candidate CV subclass table with our revised classifications. Figure 7.17 shows the location of out-of-sample CVs (middle subplot) and DNe (right subplot) when projected onto the learnt UMAP space generated without Gaia DR3 data. They are coloured in black overlaying the training set examples are colour-coded by their class label. The left subplot of the figure displays the training set examples alone for easier comparison. Figure 7.18 repeats this but for the UMAP projection generated with Gaia DR3 data. While PCA and GTM offer alternative perspectives, the clear clustering and detail provided by UMAP demonstrate this approach's efficacy.

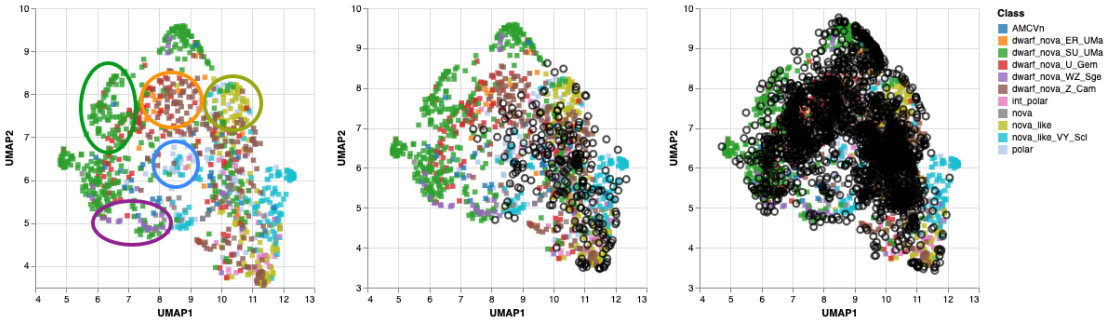


FIGURE 7.17: UMAP training set projection without Gaia DR3 data with out-of-sample example projections overlaid. The left subplot is training set projection alone, the middle subplot is the same but with out-of-sample CVs overlaid, the right subplot is the same as the first but with out-of-sample DN overlaid. Out-of-sample examples are displayed as black open circles. The left plot has been annotated to indicate regions of particular interest where out-of-sample source projections have been investigated (see text).

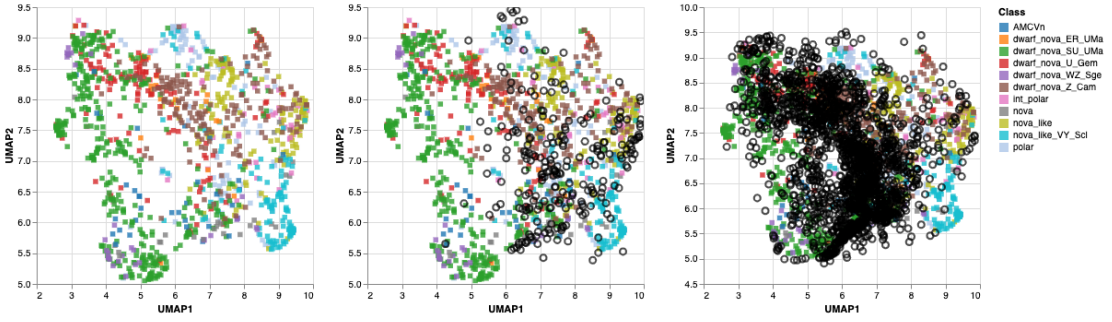


FIGURE 7.18: UMAP training set projection with Gaia DR3 data with out-of-sample example projections overlaid. The layout is the same as for Figure 7.17.

For both cases — without and with Gaia DR3 data — the out-of-sample CVs tend to be located in regions where no particular class from the training set dominates, resulting in a mixture of classes. Other key observations include: very few out-of-sample CVs are projected onto the SU UMa and U Gem regions; out-of-sample dwarf novae (DNe) overlap with out-of-sample CVs, extending into the U Gem territory and partially into SU UMa territory; and very few out-of-sample sources are located where VY Scl systems are found.

To apply a more granular class label to the out-of-sample sources, I selected several regions from the UMAP projection generated without Gaia DR3 data and examined the light curves of the out-of-sample sources located there. The regions are circled in the left subplot of Figure 7.17 and correspond to polars (blue circle), rapidly outbursting Z Cams (orange circle), eclipsing systems (olive green), WZ Sge (purple), and frequently

outbursting SU UMa with quiescent sampling (dark green). Table 7.2 lists the sources I identified from these regions as potential candidates for a more granular class, and Figure 7.19 displays the associated light curves.

Despite my efforts, many of the sources I inspected remained difficult to classify beyond the granularity provided by the AAVSO labels. Specifically, it is challenging to confirm the presence of superoutbursts in many of the out-of-sample DNe, making it hard to confidently assign them the SU UMa candidate label. Additionally, while many out-of-sample DNe in the Z Cam region exhibit rapid outbursts characteristic of the class, few show the desired standstills required for a Z Cam classification. Despite these challenges, this exercise provides an alternative method for classifying CV sources. Like the classification models discussed in earlier chapters, this method is significantly less labour-intensive than manually inspecting all light curves from an alert stream.

TABLE 7.2: List of out-of-sample sources reclassified with candidate labels based on their projection onto the lower dimensional space modelled by UMAP without the use of Gaia DR3 data. The AAVSO classifications are either CV, to designate a broad CV classification, or UG, which is an abbreviation of U Gem but is the classification tag AAVSO uses to indicate a broad dwarf nova classification. The final column represents the new granular classifications.

ZTF Object ID	RA	Dec	AAVSO Class	Revised candidate class
ZTF18abddipi	290.7380	42.0751	UG	Eclipsing
ZTF18abzyvjx	63.3718	31.2745	CV	Eclipsing
ZTF18abwkyxs	119.8600	59.8976	CV	Polar
ZTF18aaagcqv	131.0167	79.7357	UG	Polar
ZTF19aamwsgn	261.8045	32.2613	UG	Polar
ZTF18abcnnfj	338.8415	33.0455	UG	Polar
ZTF19aaaqkid	243.1549	-12.2011	CV	Polar
ZTF18acmykpd	52.5621	38.1505	UG	Polar
ZTF18acrcrte	61.8757	7.1383	CV	Polar
ZTF18acnnteg	118.6093	-0.9256	CV	Polar
ZTF17aaadqia	345.2833	-1.9679	UG	Polar
ZTF17aaaehyg	331.4905	36.0586	CV	Polar
ZTF18abolzdh	330.6129	40.2360	CV	Polar
ZTF17aadkisd	94.5388	22.1366	CV	Polar
ZTF18aayefwp	306.5040	33.6623	CV	Polar

Continuation of Table 7.2

ZTF Object ID	RA	Dec	AAVSO Class	Revised candidate class
ZTF18abydjvi	57.5447	32.5416	CV	Polar
ZTF18abmarba	320.2872	30.5707	UG	SU UMa
ZTF18aaaekuo	94.9769	9.3086	UG	SU UMa
ZTF18abxywka	331.6161	37.7817	UG	SU UMa
ZTF18aabejyh	250.9129	22.5237	UG	SU UMa
ZTF17aaaiiqk	110.9465	4.1956	UG	SU UMa
ZTF18aajpqbj	227.7908	57.6834	UG	SU UMa
ZTF19adbqznr	105.9654	32.8988	UG	WZ Sge / SU UMa
ZTF22abftmib	90.4888	39.0646	UG	WZ Sge / SU UMa
ZTF18aboslis	7.2266	43.1956	UG	WZ Sge / SU UMa
ZTF21abpvsig	345.6750	44.7156	UG	WZ Sge / SU UMa
ZTF18abmnbne	334.3819	46.9925	CV	WZ Sge / SU UMa
ZTF18abuppce	68.5023	71.4068	UG	Z Cam
ZTF18abbyacy	287.4810	-22.8004	UG	Z Cam
ZTF18abmoogg	84.5831	41.8565	UG	Z Cam
ZTF18abmrfmb	294.9945	-4.7402	UG	Z Cam
ZTF18abnpcjw	292.9549	12.7362	UG	Z Cam
ZTF18abindhr	286.6716	-14.2476	CV	Z Cam
ZTF18abmprgc	300.8650	20.0728	UG	Z Cam
ZTF18abmefkz	305.5561	19.7838	UG	Z Cam
ZTF18aaydzsl	278.5620	3.1977	UG	Z Cam

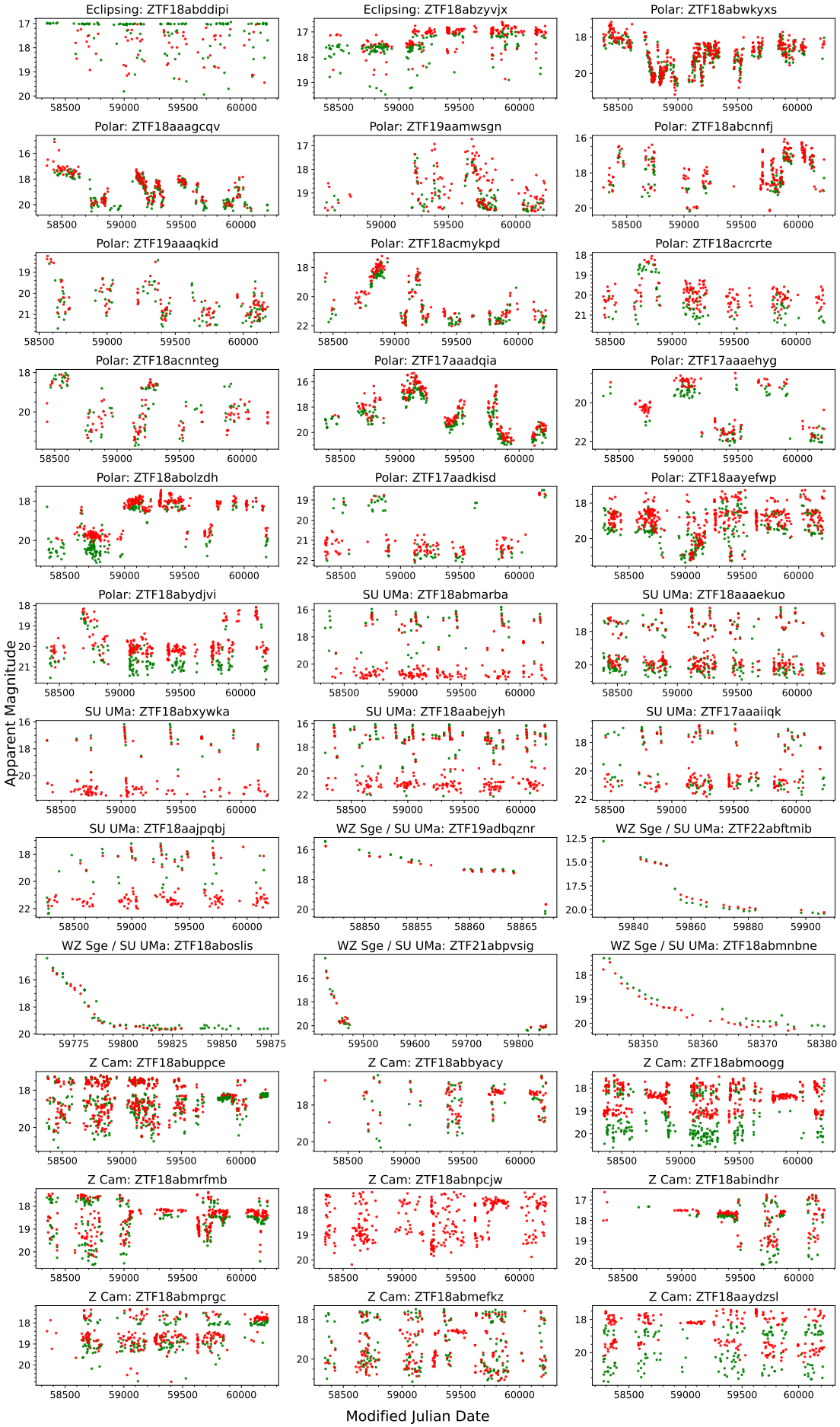


FIGURE 7.19: Light curves for a selection of sources identified in Table 7.2

7.4 Discussion

This chapter explored the use of several dimensionality reduction techniques — PCA, t-SNE, UMAP, and GTM — on ZTF light curves and their associated Gaia DR3 metadata, to examine the potential substructure within the cataclysmic variable (CV) classification scheme. The primary objectives were to explore the presence and nature of substructure, to understand the challenges in classifying CVs, and to evaluate unsupervised learning methods as an aid for CV classification (or reclassification).

7.4.1 Dimensionality Reduction Results

7.4.1.1 PCA Projections

The PCA projections revealed patterns reflecting the evolution of non-magnetic hydrogen CVs. These projections show a progression from long-period systems, like nova-likes, to short-period systems such as WZ Sge systems. However, significant overlap between different CV classes remains, and the inclusion of Gaia DR3 data only offers limited improvement in class separation. This suggests that linear methods such as PCA may not fully capture the complexity of CV variability and that non-linear techniques are necessary to uncover the expected substructures.

7.4.1.2 Non-linear Algorithms

In contrast, t-SNE and UMAP produced more distinct clusters or substructures in the data while preserving some semblance of global structure. However, substantial scatter between the clusters remained, and it was difficult to determine which algorithm yielded better separation — hyperparameter choices played a significant role. Gaia DR3 data did not drastically improve cluster separation, imputation may have affected the results, as between 10% and 30% of the Gaia-related features were missing.

CV subclass distributions were consistent across the t-SNE, UMAP, and GTM projections, irrespective of whether Gaia DR3 data was included. This consistency suggests that the patterns are not solely a result of random initialisations or hyperparameter choices, but rather reflect meaningful structure in the data.

7.4.2 Interpreting the Results

Feature projections aid in understanding the properties of examples located in specific regions of the 2D projections. However, interpretability suffers due to the overlapping of examples and scatter with them. GTM reference maps offer a much clearer and more interpretable method for analysis, with the ability to better assess feature relevance. For example, features such as parallax, proper motion, and colour-related features exhibited clear structure in the reference maps, which was not apparent in the feature projections.

The noticeable scatter between clusters in t-SNE and UMAP projections suggests that CVs may not neatly belong to discrete classes but rather form a continuum of variability types. This continuum is likely influenced by both observational limitations (e.g., survey sensitivity, sampling cadence) and evolutionary factors, which the existing CV classification framework only partially accounts for. This finding emphasises the challenges of applying supervised learning to the ZTF CV dataset and suggests that a more nuanced approach, such as unsupervised learning, may be needed to better capture the diversity of CV variability.

7.4.3 Connecting Identified Features to Physical Understanding

This investigation has helped to elucidate the features that effectively distinguish between CV subclasses. The features provide significant insight into the underlying physical processes driving their behaviour. Each feature represents a measurable property of the system, tying observational data to fundamental astrophysical phenomena.

Parallax and proper motion highlight the relationship between the intrinsic luminosity of different CV subclasses and their detectability. Colour metrics provide a window into the flux contributions from the white dwarf, accretion disk, and donor star. This informs our understanding of accretion dynamics, energy distribution, and the evolutionary states of the systems. The standard deviation of epochal colour (*clr_std*) and the distribution of flux values (*FluxPercentileRatioMid50*) are particularly useful in characterising eclipsing from non-eclipsing systems, while amplitude and periodicity measures capture unique signatures of outbursts and superoutbursts. Combined they, directly connect to orbital geometries and disk instabilities.

7.4.4 Opportunities for Further Research

The projection of out-of-sample sources onto the learnt UMAP low-dimensional space resulted in candidate sub-classifications for several sources, providing a valuable method for uncovering the nature of unclassified sources. Based on this success, the implementation of the UMAP model within our ZTF pipeline (Chapter 6) is something that can be explored.

The localisation of eclipsing CVs and polars is crucial for advancing CV research. For example, Polars, which make up around 2% of the AAVSO VSX catalogued CVs, are underrepresented in confirmed CV lists. Expanding the sample size of these systems could significantly contribute to research on magnetically controlled accretion, which has important implications for understanding the physical processes governing CVs. Furthermore, the ability to identify the features most influential in distinguishing such sources from others is useful information for conducting automated searches within astronomical databases.

7.4.5 Limitations of the Approach

Despite the success of unsupervised learning, several limitations remain. For example, the localisation of AM CVn systems in the 2D projections is hindered by their relatively small sample size and diverse variability characteristics, which could reflect ongoing evolutionary processes (e.g., changes in outburst frequency). Also, the localisation of intermediate polars remains incomplete, likely due to diverse variability characteristics. This may be a consequence of a continuum in magnetic field strengths and mass-transfer rates, leading to structures similar to those observed in non-magnetic hydrogen CVs.

Data quality issues, such as sparse or inconsistent sampling, can obscure the classification of certain CV types. Furthermore, label noise could cause class blending in the projections, requiring closer inspection and potential revision of class labels.

7.5 Conclusions

In summary, this chapter's exploration of dimensionality reduction techniques has revealed an intricate structure of CV types, emphasising the challenges in classification

due to observational constraints and evolutionary factors. The observed continuum of CV classes in the projections highlights the complexity of their evolution, suggesting that the current classification framework oversimplifies the diversity shaped by factors such as orbital period, accretion rate, and magnetic field strength. The clear localisation of eclipsing systems and other distinct CV subclasses, such as polars and SU UMa systems, provides candidates for follow-up studies. These studies provide opportunities to refine orbital and physical parameters and study specific variability mechanisms, such as magnetically controlled accretion and superhump phenomena.

Through this analysis, unsupervised learning emerges as a powerful tool to enhance our comprehension of these astrophysical phenomena and inform the broader field of CV research.

Chapter 8

Discussion and Conclusions

This investigation into the effectiveness of Machine Learning in the identification/classification of CVs and their various subtypes from within wide field transient surveys has spanned supervised and unsupervised learning techniques, and low to high cadence photometry. In discussing this journey, I summarise and highlight the significance, address limitations, and explain the implications of my research. I then look at how this research may progress.

8.1 Summary and significance

The application of ML techniques to the transient stream of Gaia Science Alerts resulted in a Random Forest model capable of distinguishing between AGN, SNe, CVs, and YSOs based on Gaia G band light curves and Gaia metadata (parallax, proper motion, BP-RP colour...). Evaluated on a test set, an F1-score of 89% for the CV class was achieved. While metadata provided improvement in model performance compared to where only light curve-derived features were used, the difference in classification performance was only 1.9% in accuracy, demonstrating the effectiveness of the feature extraction process. When applied to the list of $> 13,000$ previously unclassified targets within GSA, the model predicted 2,833 to be of the CV class. A spectroscopic investigation of a small randomly selected subset of the brightest of these sources (15 out of 220) resulted in spectroscopic confirmation of the CV class for all 15. The model demonstrated that despite the low cadence photometry (2-4 weeks), the use of ML is valuable in paring

down a large list of unclassified transients to numbers more manageable for human inspection.

The multi-band higher cadence photometry of ZTF light curves provided the opportunity to identify and distinguish between CV subclasses within the ZTF alert stream. Alerts filtering proved to be valuable too in honing in on sources most likely to belong to the CV class with both known and potential CVs retained in large part due to g-r colour thresholds. The classifier itself, built using the XGBoost algorithm trained on both g and r band light curves and Gaia DR3 metadata, achieved an AUC score of 0.92 for distinguishing between CV classes. While analysis of the GTM latent space representations of the class posterior probability space of this classifier revealed the impact that CV evolution may have on the prediction pattern. GTM also proved valuable in identifying features most relevant for the identification of individual CV classes. Implementation of the pipeline on the ZTF stream throughout June 2023 provided between 50-200 sources per night for input into the classifier, 45% of which are reported as either confirmed or candidate CVs. Despite an estimated 5-10% contamination from AGN, YSO, and variable stars, 51 intriguing and previously unreported CV candidates spanning a range of CV types, including AM CVns and polars, were found. This phase of research demonstrated the power of higher cadence and multiband photometry for delving into the identification of different CV subtypes with ML. Techniques such as GTM class maps also highlighted the impact of CV evolution on classification.

As a continuation of the evolution theme, the use of dimensionality reduction techniques on the ZTF dataset served to indicate that CV subtypes seem to form a continuum where no clear class boundaries appear to exist. This is especially evident for hydrogen-rich non-magnetic CVs for which the sample size is sufficient to see these fuzzy class boundaries. PCA is effective in viewing global structure in CV features space where a clear trend of long to short-period systems is evident. With t-SNE and UMAP, the local structure in the data becomes more evident. Whilst good separation is seen between several classes, both the impact of particulars of the survey (limiting magnitude, sampling cadence, seasonal gaps) and evolutionary factors are evident in the 2D projection substructure. The use of GTM helped to clarify the relevance of each feature through substructures in reference maps. The results highlight the challenges in applying supervised learning to our ZTF CV dataset. As a bonus, the separation of clearly eclipsing

systems from other CVs opens up the opportunity to discover more such systems when new data is projected onto the low-dimensional space with the UMAP models.

8.2 Comparison with existing literature

Existing literature on the use of Machine Learning for source classification within transient surveys has regarded CVs as a broad transient class without further subdivision. Examples include a multiclass model trained on CRTS light curves (Neira et al., 2020) achieving an F1-score of 75% for the CV class; and Sánchez-Sáez et al. (2021) achieving CV recall scores of between 61% and 72% for models trained on ZTF light curves. My research into GSA transient classification produced a 4-class model whose results compare favourably with those examples (80% in both CV F1 and recall scores). However, such comparisons do not account for differences in survey instruments, sky coverage, observing cadence, and waveband. Furthermore, comparisons do not consider differences in transient classes attempted for classification and the ML methods employed. A more relevant comparison may be made with recent work by Rimoldini et al. (2022) where the classification of 12.5 million sources into 25 classes was attempted, achieving a CV F1 score of 23.7%. However, this research differs from my GSA classification research in class structure, data input, and ML methods adopted, again making comparison difficult.

Concerning ZTF research, one may compare the CV candidates output by my pipeline with the non-ML filter-based approach of Szkody et al. (2020, 2021), where over two years, 497 new strong CV candidates were uncovered from ZTF alerts by applying simple colour, amplitude and variability timescale filters. The ZTF alerts classification pipeline of this research produces 51 new CV candidates in June 2023, which is comparable to the output of Szkody et al. (2020, 2021). Nine of the candidates from my ZTF exploration have been assigned the AM CVn classification with class probabilities between 0.26 and 0.70. However, there is as yet no spectroscopic confirmation of these nine, which means that a comparison to work by van Roestel et al. (2021) may not be made. In that work, an extension of the filter approach of Szkody et al. (2020, 2021) was performed employing Gaia and PanSTARRS colours to identify and spectroscopically confirm nine outbursting AM CVns within the whole corpus of ZTF alerts.

So far as I am aware, there is no such literature focusing on ML-based CV subclass classification or the use of unsupervised learning specifically focused on CV subclasses.

8.3 Limitations

While honing in on a population of CVs from alert streams was generally effective for classification research with GSA and ZTF, the requirements of subclass classification are such that low sample sizes of particular CV classes, label noise and inconsistent/low sampling cadence impacted the classification performance.

The adverse impact of low sample size is particularly evident for the intermediate polar class. Whilst synthetic samples from the feature space were generated to address the class imbalance, the examples in this class displayed differences in terms of the degree of long-term variability (weeks to months) as well as the presence or absence of dwarf nova outbursts. Consequently, their location in feature space is not localised (as demonstrated with dimensionality reduction techniques). Furthermore, synthetic samples generated with ADASYN, through feature space interpolation, may have only served to exacerbate classification performance issues.

Classifying CVs into subclasses is a challenge, even for experts in the field, due to their complex variability, subjective interpretation of features, and incomplete data, which often leaves key characteristics unobserved. Furthermore, the boundaries between subclasses are often indistinct, creating overlap and complicating clear categorisation. Despite efforts to reduce label noise in the ZTF research, these factors introduce classification uncertainties that are difficult to quantify. The exploration of unsupervised learning with the ZTF dataset brings these challenges into focus. Not only are class boundaries poorly defined, but examples with few data points in their light curves tend to pool together into systems with a mixture of CV subclasses.

To address these challenges, adopting stricter inclusion criteria — focusing on well-sampled light curves and clear examples of each class — can help define class boundaries more effectively, simplifying the learning process. However, this approach comes at the expense of a reduced dataset size. A cost that may be mitigated with data augmentation. The resultant classification of unseen examples would then focus less on definitive class labels and more on the probability of class belonging. This, however, does not address

the issue of the diverse variability of low sample size classes, such as intermediate polars. Adoption of a phenomenological classification approach, as in van Roestel et al. (2021), where types of variability are the focus of classification could form another approach, though due to multiple phenomena existing in any given light curves, this would be handled under the multi-label paradigm, where multiple nonexclusive labels may be assigned to each instance (Hastie et al., 2003).

8.4 Implications

My research has attempted to address a gap in ML-based time domain source classification, that of CV subclass classification. The effectiveness of the ZTF alerts pipeline drastically reduces the requirement for human inspection which is a significant feature of works by Szkody et al. (2020), Szkody et al. (2021), and van Roestel et al. (2021). The aspect of CV subclass classification provides greater depth by focusing on the CV substructure rather than treating this diverse transient class as one group. This makes it easier for research groups focused on particular CV subclasses to identify their sources of interest.

In Chapter 6, the imprint of CV evolution on classification was alluded to, though made more explicit with the use of dimensionality reduction techniques in Chapter 7. Rather than classification into types with fuzzy class boundaries, the projection of unseen data onto the 2D space generated by PCA, UMAP and GTM, now provides an alternative route to understanding the nature of new CV candidates.

In the coming years, the discovery potential of CVs will be further enhanced by the advent of new transient surveys resulting in a large increase in the rate at which new time-varying sources are discovered. Leading the way in this respect will be the Vera Rubin Observatory Legacy Survey of Space and Time (LSST; Ivezić et al. 2019), generating up to 10^7 alerts per night (Matheson et al., 2021). The research conducted has demonstrated the success of ML applications to CV subclass identification/classification, and such a pipeline implemented on these surveys will be vital if we are to distinguish the rare varieties of CV from the deluge of transient phenomena.

8.5 Future research directions

8.5.1 Extension to other surveys

The research provided the first steps into ML-based subclass identification/classification of CVs. The research was confined to the Gaia and ZTF surveys. Exploration of and/or inclusion of data from multiple surveys would complement the research and provide a greater understanding of the challenges of such granular classification attempts. One example would be the higher cadence surveys such as ASAS-SN, with nightly observation or the Transiting Exoplanet Survey Satellite (TESS; Ricker et al. 2015) with survey cadences of order minutes. Such an exploration may well reveal a detailed substructure where unsupervised learning is used.

8.5.2 Alternative representations

This research has prioritised the use of light curve-derived hand-crafted features—such as statistical, periodicity-based, and percentile-based metrics—to ensure interpretability. However, incorporating alternative representations, such as those based on computer vision techniques, presents a promising direction. For instance, van Roestel et al. (2021) explored this approach in the ZTF Source Classification Project, where they classified broad transient classes using a dmdt representation of ZTF light curves combined with traditional hand-crafted features within a Convolutional Neural Network (CNN). The CNN's convolutional layers processed the dmdt image representation of the light curves, while the hand-crafted features were appended to the final convolutional layer's output before entering the fully connected layers. While their results were comparable to models using only hand-crafted features, fine-tuning the dmdt histogram binning could have potentially improved performance.

In my experimentation with ZTF data, I explored dmdt mapping specifically for CV subclasses. This preliminary effort did not involve optimising bin settings or integrating additional features. Sample dmdt representations and the corresponding confusion matrix are shown in Figures 8.1 and 8.2. The visual distinctions in dmdt mappings among CV subclasses are encouraging, indicating the potential for subclass differentiation. However, intra-class variability in these mappings remains unexplored. The

confusion matrix highlights classification patterns similar to those observed in Chapter 6, with the strongest performance for dwarf nova classes, nova-likes, VY Scl, and polars. Class-specific precision ranges from 0.19 to 0.85, recall from 0.31 to 0.63, and F1-scores from 0.27 to 0.73. These results underscore the potential of dmdt mapping as a complementary tool for CV classification.

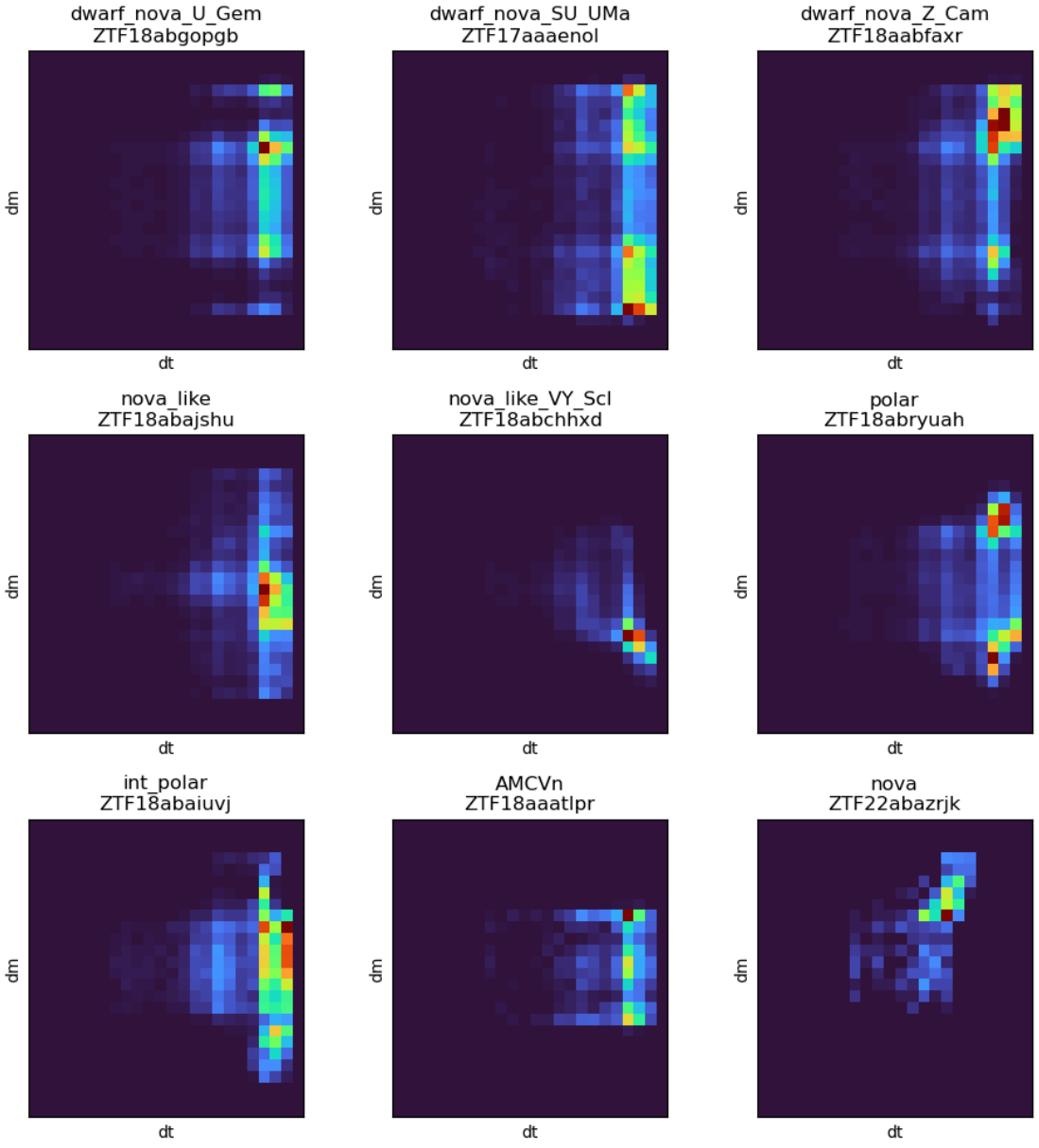


FIGURE 8.1: The dmdt representations of a member of each of the CV subclasses. The dm bins span the vertical axis, while the dt bins span the horizontal axis.

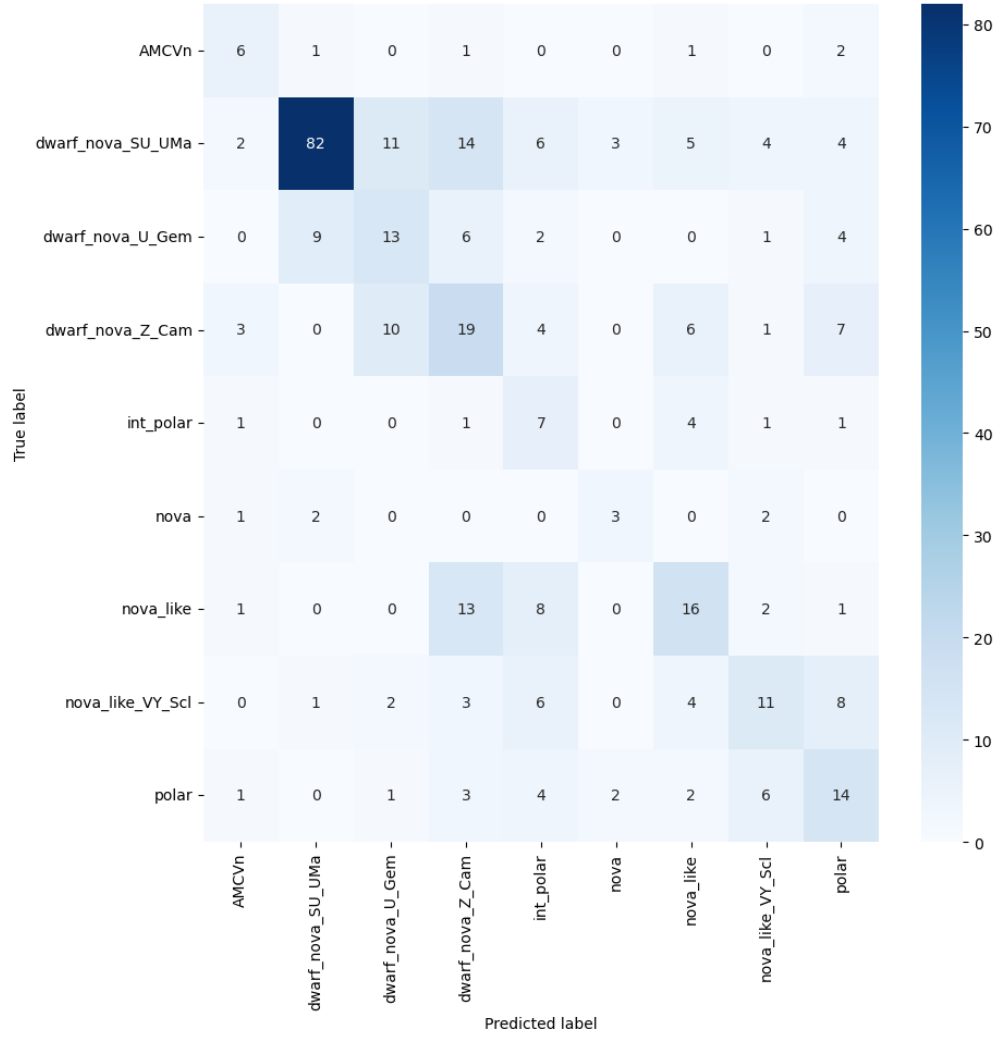


FIGURE 8.2: The confusion matrix of the CNN model trained on dmdt representations of the CVs in the ZTF light curve dataset filtered such that only those sources with a g band light curve with 20 points or more are included.

8.6 Conclusions

In conclusion, this thesis has made significant contributions to the classification of cataclysmic variables (CVs) through the application of machine learning (ML) techniques to wide-field transient surveys. By automating the identification and subclassification of CVs, the reliance on human inspection has been greatly reduced. The models developed, including a Random Forest classifier for Gaia Science Alerts and an XGBoost classifier for ZTF alerts, have outperformed previous attempts, achieving high F1-scores in CV classification. Despite challenges such as low sample sizes and label noise, the research has provided valuable insights into the substructure of CVs.

Looking ahead, future research will focus on extending this analysis to other surveys, such as the Vera Rubin Observatory's Legacy Survey of Space and Time (LSST), which promises to revolutionize transient surveys with its vast data volume. Additionally, exploring alternative representations, including computer vision techniques, could further enhance classification accuracy. This work establishes a strong foundation for continued advancements in ML-based CV classification and offers promising avenues for deeper insights into these fascinating astrophysical phenomena.

Bibliography

Analytics Vidhya, 2021, Convolutional Neural Networks (CNN), https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/

Anzolin G., de Martino D., Bonnet-Bidaud J.-M., Mouchet M., Gänsicke B. T., Matt G., Mukai K., 2008, Astronomy and Astrophysics, 489, 1243

Bellm E. C., Kulkarni S. R., Graham M. J., Dekany R., Smith R. M., Riddle R., Masci F. J., Helou G., Prince T. A., Adams S. M., et al., 2019, Publications of the Astronomical Society of the Pacific, 131, 018002

Bellman R., 1957, Science, 153, 34

Bishop C. M., Svensén M., Williams C. K. I., 1998, Neural Computation, 10, 215

Blagorodnova N., Neill J. D., Walters R., Kulkarni S. R., Fremling C., Ben-Ami S., Dekany R. G., Fucik J. R., Konidaris N., Nash R., et al., 2018, Publications of the Astronomical Society of the Pacific, 130, 035003

Bode M. F., Evans A., 2008, Classical Novae. Vol. 43, Cambridge University Press

Breedt E., Gänsicke B. T., Drake A. J., Rodríguez-Gil P., Parsons S. G., Marsh T. R., Szkody P., Schreiber M. R., Djorgovski S. G., 2014, Monthly Notices of the Royal Astronomical Society, 443, 3174

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

Brockwell P. J., Davis R. A., 2002, Introduction to time series and forecasting, 2nd edn. Springer texts in statistics, Springer, New York

Buat-Ménard V., Hameury J. M., Lasota J. P., 2001a, Astronomy and Astrophysics, 366, 612

Buat-Ménard V., Hameury J. M., Lasota J. P., 2001b, Astronomy and Astrophysics, 369, 925

Burlak M. A., Henden A. A., 2008, Astronomy Letters, 34, 241

Cabral J. B., Sánchez B., Ramos F., Gurovich S., Granitto P. M., Vanderplas J., 2018, Astronomy and Computing, 25, 213

Campbell H. C., Marsh T. R., Fraser M., Hodgkin S. T., de Miguel E., Gänsicke B. T., Steeghs D., Hourihane A., Breedt E., Littlefair S. P., et al., 2015, Monthly Notices of the Royal Astronomical Society, 452, 1060

Cannizzo J. K., Nelemans G., 2015, The Astrophysical Journal, 803, 19

Cao Y., Nugent P. E., Kasliwal M. M., 2016, Publications of the Astronomical Society of the Pacific, 128, 114502

Carrasco-Davis R., Reyes E., Valenzuela C., Förster F., Estévez P. A., Pignata G., Bauer F. E., Reyes I., Sánchez-Sáez P., Cabrera-Vives G., et al., 2021, The Astronomical Journal, 162, 231

Carroll B. W., Ostlie D. A., 1996, An Introduction to Modern Astrophysics. Cambridge University Press

Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P., 2002, Journal of artificial intelligence research, 16, 321

Chen C., Breiman L., 2004, University of California, Berkeley

Chen T., Guestrin C., , 2016, XGBoost: A Scalable Tree Boosting System

Chollet F., 2021, Deep Learning with Python, Second Edition. Manning

Chomiuk L., Metzger B. D., Shen K. J., , 2020, New Insights into Classical Novae

Cortes C., Vapnik V., 1995, Machine Learning, 20, 273

Covington A. E., Shaw A. W., Mukai K., Littlefield C., Heinke C. O., Plotkin R. M., Barrett D., Boardman J., Boyd D., Brincat S. M., et al., 2022, The Astrophysical Journal, 928, 164

Cropper M., 1990, Space Science Reviews, 54, 195

Cui X.-Q., Zhao Y.-H., Chu Y.-Q., Li G.-P., Li Q., Zhang L.-P., Su H.-J., Yao Z.-Q., Wang Y.-N., Xing X.-Z., Li X.-N., Zhu Y.-T., Wang G., Gu B.-Z., Luo A.-L., Xu X.-Q., Zhang Z.-C., Liu G.-R., Zhang H.-T., Yang D.-H., Cao S.-Y., Chen H.-Y., Chen J.-J., Chen K.-X., Chen Y., Chu J.-R., Feng L., Gong X.-F., Hou Y.-H., Hu H.-Z., Hu N.-S., Hu Z.-W., Jia L., Jiang F.-H., Jiang X., Jiang Z.-B., Jin G., Li A.-H., Li Y., Li Y.-P., Liu G.-Q., Liu Z.-G., Lu W.-Z., Mao Y.-D., Men L., Qi Y.-J., Qi Z.-X., Shi H.-M., Tang Z.-H., Tao Q.-S., Wang D.-Q., Wang D., Wang G.-M., Wang H., Wang J.-N., Wang J., Wang J.-L., Wang J.-P., Wang L., Wang S.-Q., Wang Y., Wang Y.-F., Xu L.-Z., Xu Y., Yang S.-H., Yu Y., Yuan H., Yuan X.-Y., Zhai C., Zhang J., Zhang Y.-X., Zhang Y., Zhao M., Zhou F., Zhou G.-H., Zhu J., Zou S.-C., 2012, Research in Astronomy and Astrophysics, 12, 1197

Darnley M. J., Bode M. F., Kerins E., Newsam A. M., An J., Baillon P., Belokurov V., Novati S. C., Carr B. J., Crézé M., et al., 2006, Monthly Notices of the Royal Astronomical Society, 369, 257

Darnley M. J., Henze M., 2020, Advances in Space Research, 66, 1147

Darnley M. J., Henze M., Bode M. F., Hachisu I., Hernanz M., Hornoch K., Hounsell R., Kato M., Ness J. U., Osborne J. P., et al., 2016, The Astrophysical Journal, 833, 149

Darnley M. J., Henze M., Steele I. A., Bode M. F., Ribeiro V. A. R. M., Rodríguez-Gil P., Shafter A. W., Williams S. C., Baer D., Hachisu I., et al., 2015, Astronomy and Astrophysics, 580, A45

Darnley M. J., Ribeiro V. A. R. M., Bode M. F., Hounsell R. A., Williams R. P., 2012, The Astrophysical Journal, 746, 61

de Martino D., Bernardini F., Mukai K., Falanga M., Masetti N., 2020, Advances in Space Research, 66, 1209

Demircioğlu A., 2021, Insights into Imaging, 12, 172

Dimple Misra K., Arun K. G., 2023, The Astrophysical Journal Letters, 949, L22

Downes R., Webbink R. F., Shara M. M., 1997, Publications of the Astronomical Society of the Pacific, 109, 345

Downes R. A., Shara M. M., 1993, Publications of the Astronomical Society of the Pacific, 105, 127

Downes R. A., Webbink R. F., Shara M. M., Ritter H., Kolb U., Duerbeck H. W., 2001, Publications of the Astronomical Society of the Pacific, 113, 764

Drake A. J., Djorgovski S. G., Mahabal A., Beshore E., Larson S., Graham M. J., Williams R., Christensen E., Catelan M., Boattini A., et al., 2009, The Astrophysical Journal, 696, 870

Drake A. J., Gänsicke B. T., Djorgovski S. G., Wils P., Mahabal A. A., Graham M. J., Yang T.-C., Williams R., Catelan M., Prieto J. L., et al., 2014, Monthly Notices of the Royal Astronomical Society, 441, 1186

Dubus G., Otulakowska-Hypka M., Lasota J.-P., 2018, Astronomy & Astrophysics, 617, A26

Duffy C., Ramsay G., Steeghs D., Dhillon V., Kennedy M. R., Mata Sánchez D., Ackley K., Dyer M., Lyman J., Ulaczyk K., et al., 2021, Monthly Notices of the Royal Astronomical Society, 502, 4953

Duffy C., Ramsay G., Wu K., Mason P. A., Hakala P., Steeghs D., Wood M. A., 2022, Monthly Notices of the Royal Astronomical Society, 516, 3144

Eggleton P. P., 1983, The Astrophysical Journal, 268, 368

Faulkner J., Flannery B. P., Warner B., 1972, The Astrophysical Journal, 175, L79

Fernández A., García S., Galar M., Prati R. C., Krawczyk B., Herrera F., 2018, Learning from Imbalanced Data Sets, first edn. Springer International Publishing, Springer Cham

Ferrario L., de Martino D., Gänsicke B. T., 2015, Space Science Reviews, 191, 111

Flesch E. W., , 2019, The Million Quasars (Milliquas) Catalogue, v6.4

Frank J., King A., Raine D. J., 2002, Accretion Power in Astrophysics: Third Edition. Cambridge University Press

Frenay B., Verleysen M., 2014, IEEE Transactions on Neural Networks and Learning Systems, 25, 845

Freund Y., Schapire R. E., 1997, Journal of Computer and System Sciences, 55, 119

Förster F., Cabrera-Vives G., Castillo-Navarrete E., Estévez P. A., Sánchez-Sáez P., Arredondo J., Bauer F. E., Carrasco-Davis R., Catelan M., Elorrieta F., et al., 2021, The Astronomical Journal, 161, 242

Gaia-Collaboration Prusti T., de Bruijne J. H. J., Brown A. G. A., Vallenari A., Babusiaux C., Bailer-Jones C. A. L., Bastian U., Biermann M., Evans D. W., et al., 2016, Astronomy and Astrophysics, 595, A1

Gaia-Collaboration Vallenari A., Brown A., Prusti T., et al. 2022, Astronomy & Astrophysics

Garraffo C., Drake J. J., Alvarado-Gomez J. D., Moschou S. P., Cohen O., 2018, The Astrophysical Journal, 868, 60

Gaspar H. A., 2018, Journal of Open Research Software

Georganti M., Knigge C., Castro Segura N., Long K. S., 2022, Monthly Notices of the Royal Astronomical Society, 511, 5385

Giovannelli F., 2008, Chinese Journal of Astronomy and Astrophysics Supplement, 8, 237

Goldstein A., Veres P., Burns E., Briggs M., Hamburg R., Kocevski D., Wilson-Hodge C., Preece R., Poolakkil S., Roberts O., 2017, The Astrophysical Journal Letters, 848, L14

Goldstein D. A., D'Andrea C. B., Fischer J. A., Foley R. J., Gupta R. R., Kessler R., Kim A. G., Nichol R. C., Nugent P. E., Papadopoulos A., et al., 2015, The Astronomical Journal, 150, 82

Green M. J., Hermes J. J., Marsh T. R., Steeghs D. T. H., Bell K. J., Littlefair S. P., Parsons S. G., Dennihy E., Fuchs J. T., Reding J. S., et al., 2018, Monthly Notices of the Royal Astronomical Society, 477, 5646

Gänsicke B. T., Dillon M., Southworth J., Thorstensen J. R., Rodríguez-Gil P., Aungwerojwit A., Marsh T. R., Szkody P., Barros S. C. C., Casares J., et al., 2009, Monthly Notices of the Royal Astronomical Society, 397, 2170

Hachisu I., Kato M., Kiyota S., Kubotera K., Maehara H., Nakajima K., Ishii Y., Kamada M., Mizoguchi S., Nishiyama S., et al., 2006, The Astrophysical Journal, 651, L141

Haibo H., Yang B., Garcia E. A., Shutao L., 2008, in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence) Adasyn: Adaptive synthetic sampling approach for imbalanced learning. pp 1322–1328

Hameury J. M., 2020, Advances in Space Research, 66, 1004

Hameury J. M., Lasota J. P., 2017, Astronomy and Astrophysics, 602, A102

Hameury J. M., Lasota J. P., Shaw A. W., 2022, Astronomy and Astrophysics, 664, A7

Harrington R. G., 1952, Publications of the Astronomical Society of the Pacific, 64, 275

Hastie T. J., Tibshirani R., Friedman J. H., 2003, in Springer Series in Statistics The elements of statistical learning: Data mining, inference, and prediction, 2nd edition

Hawley J. F., Balbus S. A., 1998, in Howell S., Kuulkers E., Woodward C., eds, Wild Stars in the Old West Vol. 137, Anomalous viscosity in accretion disks. p. 273

Hellier C., 1993, Monthly Notices of the Royal Astronomical Society, 265, L35

Hellier C., 2001, Cataclysmic Variable Stars. Springer London

Hellier C., Beardmore A. P., 2002, Monthly Notices of the Royal Astronomical Society, 331, 407

Henze M., Pietsch W., Haberl F., Della Valle M., Sala G., Hatzidimitriou D., Hofmann F., Hernanz M., Hartmann D. H., Greiner J., 2014, Astronomy and Astrophysics, 563, A2

Hessman F. V., Gänsicke B. T., Mattei J. A., 2000, Astronomy and Astrophysics, 361, 952

Hillman Y., Prialnik D., Kovetz A., Shara M. M., 2015, Monthly Notices of the Royal Astronomical Society, 446, 1924

Hillman Y., Prialnik D., Kovetz A., Shara M. M., 2016, The Astrophysical Journal, 819, 168

Hjorth J., Bloom J. S., 2012, Cambridge Astrophysics Series, 51, 169

Hodgkin S. T., Harrison D. L., Breedt E., Wevers T., Rixon G., Delgado A., Yoldas A., Kostrzewa-Rutkowska Z., Wyrzykowski L., van Leeuwen M., et al., 2021, Astronomy and Astrophysics, 652, A76

Honeycutt R. K., Kafka S., 2004, The Astronomical Journal, 128, 1279

Hope D. J. E., Copperwheat C. M., 2019, Research Notes of the American Astronomical Society, 3, 72

Hou W., Luo A. l., Li Y.-B., Qin L., 2020, The Astronomical Journal, 159, 43

Hounsell R., Darnley M. J., Bode M. F., Harman D. J., Surina F., Starrfield S., Holdsworth D. L., Bewsher D., Hick P. P., Jackson B. V., et al., 2016, The Astrophysical Journal, 820, 104

Howell S. B., Nelson L. A., Rappaport S., 2001, The Astrophysical Journal, 550, 897

Hu Z., Chen J., Jiang B., Wang W., 2021, Universe, 7, 438

Iben Icko J., Tutukov A. V., 1987, The Astrophysical Journal, 313, 727

Ivezić Z., Kahn S. M., Tyson J. A., Abel B., Acosta E., Allsman R., Alonso D., AlSayyad Y., Anderson S. F., Andrew J., et al., 2019, The Astrophysical Journal, 873, 111

Jayasinghe T., Stanek K. Z., Kochanek C. S., Way Z., Vallely P., Basinger C., Thompson T. A., Shappee B. J., Holoien T. W. S., Prieto J. L., et al., 2020, The Astronomer's Telegram, 13824, 1

Jha S. W., Maguire K., Sullivan M., 2019, Nature Astronomy, 3, 706

Jose J., 2016, Stellar Explosions: Hydrodynamics and Nucleosynthesis. CRC Press

Jäger S., Allhorn A., Bießmann F., 2021, Frontiers in big data, 4, 693674

Kalomeni B., 2012, Monthly Notices of the Royal Astronomical Society, 422, 1601

Kato M., Hachisu I., 2012, Bulletin of the Astronomical Society of India, 40, 393

Kato M., Hachisu I., 2020, Publications of the Astronomical Society of Japan, 72, 82

Kato M., Saio H., Hachisu I., 2015, The Astrophysical Journal, 808, 52

Kato T., 2015, Publications of the Astronomical Society of Japan, 67

Kato T., Kojiguchi N., 2021, Publications of the Astronomical Society of Japan, 73, 1375

Kato T., Maeda Y., Moriyama M., , 2023, Genuine standstill in the AM CVn star CR Boo

Kato T., Nogami D., Baba H., Masuda S., Matsumoto K., Kunjaya C., , 2013, Observation of ER UMa Stars

Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y., 2017, in Neural Information Processing Systems Lightgbm: A highly efficient gradient boosting decision tree

Kerins E., Darnley M. J., Duke J. P., Gould A., Han C., Newsam A., Park B. G., Street R., 2010, Monthly Notices of the Royal Astronomical Society, 409, 247

Khan F., Khan K., Singh S., 2018, Journal of Physics: Conference Series, 1060, 012014

King A. R., Cannizzo J. K., 1998, The Astrophysical Journal, 499, 348

Knigge C., 2006, Monthly Notices of the Royal Astronomical Society, 373, 484

Knigge C., Baraffe I., Patterson J., 2011, The Astrophysical Journal Supplement Series, 194, 28

Kolb U., Baraffe I., 1999, Monthly Notices of the Royal Astronomical Society, 309, 1034

Kotko I., Lasota J. P., Dubus G., Hameury J. M., 2012, Astronomy and Astrophysics, 544, A13

Krautter J., 2008, in ASP Conference Series Vol. 401, The super-soft phase in novae. p. 139

Kruse R., Mostaghim S., Borgelt C., Braune C., Steinbrecher M., 2022, Multi-layer Perceptrons. Springer International Publishing, Cham, pp 53–124

Kwak N., 2016, Introduction to Convolutional Neural Networks (CNNs)

Lamers H. J. G. L. M., Cassinelli J. P., 1999, The effects of mass loss on stellar evolution. Cambridge University Press, Cambridge, pp 385–397

Leach R., Hessman F. V., King A. R., Stehle R., Mattei J., 1999, Monthly Notices of the Royal Astronomical Society, 305, 225

LeCun Y., Bengio Y., Hinton G., 2015, Nature, 521, 436

Levitan D., Groot P. J., Prince T. A., Kulkarni S. R., Laher R., Ofek E. O., Sesar B., Surace J., 2015, Monthly Notices of the Royal Astronomical Society, 446, 391

Lindegren L., Klioner S. A., Hernández J., Bombrun A., Ramos-Lerate M., Steidelmüller H., Bastian U., Biermann M., de Torres A., Gerlach E., et al., 2021, Astronomy and Astrophysics, 649, A2

Livio M., Pringle J. E., 1994, The Astrophysical Journal, 427, 956

Lü H.-J., Yuan H.-Y., Yi T.-F., Wang X.-G., Hu Y.-D., Yuan Y., Rice J., Wang J.-G., Cao J.-X., Kong D.-F., 2022, The Astrophysical Journal Letters, 931, L23

Maccarone T. J., Kupfer T., Najera Casarrubias E., Rivera Sandoval L., Shaw A., Britt C., van Roestel J., Zurek D., , 2023, Strongly magnetized accretion in ultracompact binary systems

McInnes L., Healy J., Melville J., 2018, ArXiv, p. arXiv:1802.03426

Mahabal A. A., Sheth K., Gieseke F., Pai A., Djorgovski S. G., Drake A. J., Graham M. J., 2017, 2017 IEEE Symposium Series on Computational Intelligence (SSCI), pp 1–8

Matheson T., Stubens C., Wolf N., Lee C.-H., Narayan G., Saha A., Scott A., Soraisam M., Bolton A. S., Hauger B., et al., 2021, The Astronomical Journal, 161, 107

Meyer F., Meyer-Hofmeister E., 1983, Astronomy and Astrophysics, 121, 29

Meyer F., Meyer-Hofmeister E., 1984, Astronomy and Astrophysics, 132, 143

Mistry D., Copperwheat C. M., Darnley M. J., Olier I., 2022, Monthly Notices of the Royal Astronomical Society, 517, 3362

Mistry D., Copperwheat C. M., Darnley M. J., Olier I., 2023, Monthly Notices of the Royal Astronomical Society, 527, 8633

Muhammad Ali P., Faraj R., 2014, Data Normalization and Standardization: A Technical Report. Machine Learning Lab. Koya University

Mukai K., 2017, Publications of the Astronomical Society of the Pacific, 129, 062001

Munari U., 2012, Journal of the American Association of Variable Star Observers (JAAVSO), 40, 582

Narayan G., Zaidi T., Soraisam M. D., Wang Z., Lochner M., Matheson T., Saha A., Yang S., Zhao Z., Kececioglu J., et al., 2018, The Astrophysical Journal Supplement Series, 236, 9

Neira M., Gómez C., Suárez-Pérez J. F., Gómez D. A., Reyes J. P., Hoyos M. H., Arbeláez P., Forero-Romero J. E., 2020, The Astrophysical Journal Supplement Series, 250, 11

Nelemans G., 2005, in ASP Conference Series Vol. 330, Am cvn stars. p. 27

Nelemans G., Yungelson L. R., Portegies Zwart S. F., 2004, Monthly Notices of the Royal Astronomical Society, 349, 181

Nun I., Protopapas P., Sim B., Zhu M., Dave R., Castro N., Pichara K., , 2015, FATS: Feature Analysis for Time Series

Osaki Y., 1974, Publications of the Astronomical Society of Japan, 26, 429

Osaki Y., 1989, Publications of the Astronomical Society of Japan, 41, 1005

Osaki Y., Kato T., 2013, Publications of the Astronomical Society of Japan, 65

Otulakowska-Hypka M., Olech A., Patterson J., 2016, Monthly Notices of the Royal Astronomical Society, 460, 2526

Paczynski B., 1986, Astrophysical Journal, Part 2-Letters to the Editor (ISSN 0004-637X), vol. 308, Sept. 15, 1986, p. L43-L46., 308, L43

Paczynski B., Sienkiewicz R., 1981, The Astrophysical Journal, 248, L27

Paczynski B., Sienkiewicz R., 1983, The Astrophysical Journal, 268, 825

Paczyński B., 1967, Acta Astronomica, 17, 287

Paczyński B., 1971, Annual Review of Astronomy and Astrophysics, 9, 183

Pala A. F., Gänsicke B. T., Belloni D., Parsons S. G., Marsh T. R., Schreiber M. R., Breedt E., Knigge C., Sion E. M., Szkody P., et al., 2021, Monthly Notices of the Royal Astronomical Society, 510, 6110

Patterson J., 1984, The Astrophysical Journal Supplement Series, 54, 443

Patterson J., 1994, Publications of the Astronomical Society of the Pacific, 106, 209

Patterson J., Kemp J., Jensen L., Vanmunster T., Skillman D. R., Martin B., Fried R., Thorstensen J. R., 2000, Publications of the Astronomical Society of the Pacific, 112, 1567

Patterson M. T., Bellm E. C., Rusholme B., Masci F. J., Juric M., Krughoff K. S., Golkhou V. Z., Graham M. J., Kulkarni S. R., Helou G., et al., 2019, Publications of the Astronomical Society of the Pacific, 131, 018001

Payne-Gaposchkin C., 1964, The galactic novae. Dover Publications

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., 2011, the Journal of machine Learning research, 12, 2825

Perley D. A., Fremling C., Sollerman J., Miller A. A., Dahiwale A. S., Sharma Y., Bellm E. C., Biswas R., Brink T. G., Bruch R. J., et al., 2020, The Astrophysical Journal, 904, 35

Phillips M. M., 1993, The Astrophysical Journal, 413, L105

Piascik A. S., Steele I. A., Bates S. D., Mottram C. J., Smith R. J., Barnsley R. M., Bolton B., 2014, in Proc. SPIE Vol. 9147, Sprat: Spectrograph for the rapid acquisition of transients. p. 91478H

Podsiadlowski P., Han Z., Rappaport S., 2003, Monthly Notices of the Royal Astronomical Society, 340, 1214

Prieto J. L., Hainline K., Hickox R., Goulding A., Campillay A., Gonzalez C., Hsiao E., Shappee B., Kochanek C. S., Stanek K. Z., et al., 2013, The Astronomer's Telegram, 4999, 1

Pérez E., Zingaretti L., 2019, Genes, 10, 553

Quirk T. J., 2012, One-Way Analysis of Variance (ANOVA). Springer New York, New York, NY, pp 163–179

Ramsay G., Green M. J., Marsh T. R., Kupfer T., Breedt E., Korol V., Groot P. J., Knigge C., Nelemans G., Steeghs D., et al., 2018, Astronomy & Astrophysics, 620, A141

Ramsay G., Wheatley P. J., Norton A. J., Hakala P., Baskill D., 2008, Monthly Notices of the Royal Astronomical Society, 387, 1157

Richards J. W., Starr D. L., Butler N. R., Bloom J. S., Brewer J. M., Crellin-Quick A., Higgins J., Kennedy R., Rischard M., 2011, The Astrophysical Journal, 733, 10

Ricker G. R., Winn J. N., Vanderspek R., Latham D. W., Bakos G. A., Bean J. L., Berta-Thompson Z. K., Brown T. M., Buchhave L., Butler N. R., et al., 2015, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 014003

Riello M., De Angeli F., Evans D. W., Montegriffo P., Carrasco J. M., Busso G., Palaversa L., Burgess P. W., Diener C., Davidson M., et al., 2021, Astronomy and Astrophysics, 649, A3

Rimoldini L., Holl B., Gavras P., Audard M., De Ridder J., Mowlavi N., Nienartowicz K., Jevardat de Fombelle G., Lecoeur-Taïbi I., Karbevska L., et al.,, 2022, Gaia Data Release 3: All-sky classification of 12.4 million variable sources into 25 classes

Ritter H., Kolb U., 2003, Astronomy and Astrophysics, 404, 301

Rivera Sandoval L. E., Heinke C. O., Hameury J. M., Cavecchi Y., Vanmunster T., Tordai T., Romanov F. D., 2022, The Astrophysical Journal, 926, 10

Roelofs G. H. A., Rau A., Marsh T. R., Steeghs D., Groot P. J., Nelemans G., 2010, The Astrophysical Journal, 711, L138

Rokach L., Maimon O., 2008, Data mining with decision trees. Theory and applications. Vol. 69, World Scientific

Rosen S. R., Mason K. O., Cordova F. A., 1988, Monthly Notices of the Royal Astronomical Society, 231, 549

Ross B. C., 2014, PLOS ONE, 9, e87357

Rossi A., Rothberg B., Palazzi E., Kann D., D'Avanzo P., Amati L., Klose S., Perego A., Pian E., Guidorzi C., 2022, The Astrophysical Journal, 932, 1

Savonije G. J., de Kool M., van den Heuvel E. P. J., 1986, Astronomy and Astrophysics, 155, 51

Shafter A. W., 1992, The Astrophysical Journal, 394, 268

Shafter A. W., 1997, The Astrophysical Journal, 487, 226

Shafter A. W., 2017, The Astrophysical Journal, 834, 196

Shakura N. I., Sunyaev R. A., 1973, Astronomy and Astrophysics, 24, 337

Shappee B., Prieto J., Stanek K. Z., Kochanek C. S., Holoien T., Jencson J., Basu U., Beacom J. F., Szczygiel D., Pojmanski G., et al.,, 2014, All Sky Automated Survey for SuperNovae (ASAS-SN or "Assassin")

Sherstinsky A., 2020, Physica D Nonlinear Phenomena, 404, 132306

Simonsen M., Boyd D., Goff W., Krajci T., Menzies K., Otero S., Padovan S., Poyner G., Roe J., Sabo R., et al., 2014, Journal of the American Association of Variable Star Observers (JAAVSO), 42, 177

Singhi S. K., Liu H., , 2006, Feature subset selection bias for classification learning

Smak J., 2010, Acta Astronomica, 60, 357

Smartt S. J., Valenti S., Fraser M., Inserra C., Young D. R., Sullivan M., Pastorello A., Benetti S., Gal-Yam A., Knapic C., et al., 2015, Astronomy and Astrophysics, 579, A40

Smith K. W., Smartt S. J., Young D. R., Tonry J. L., Denneau L., Flewelling H., Heinze A. N., Weiland H. J., Stalder B., Rest A., et al., 2020, Publications of the Astronomical Society of the Pacific, 132, 085002

Smith K. W., Williams R. D., Young D. R., Ibsen A., Smartt S. J., Lawrence A., Morris D., Voutsinas S., Nicholl M., 2019, Research Notes of the AAS, 3, 26

Soley-Bori M., 2013, Dealing with missing data: key assumptions and methods for applied analysis. Boston University

Solheim J. E., 2010, Publications of the Astronomical Society of the Pacific, 122, 1133

Spruit H. C., Ritter H., 1983, Astronomy and Astrophysics, 124, 267

Spruit H. C., Ronald E. T., 1993, The Astrophysical Journal, 402, 593

Starrfield S., Iliadis C., Hix W. R., 2016, Publications of the Astronomical Society of the Pacific, 128, 051001

Steele I., Smith R., Rees P., Baker I., Bates S., Bode M., Bowman M., Carter D., Etherton J., Ford M., et al., 2004, The Liverpool Telescope: performance and first results. Vol. 5489 of SPIE Astronomical Telescopes + Instrumentation, SPIE

Stetson P. B., 1996, Publications of the Astronomical Society of the Pacific, 108, 851

Strope R. J., Schaefer B. E., Henden A. A., 2010, The Astronomical Journal, 140, 34

Sun Y., Cheng Z., Ye S., Ding R., Peng Y., Zhang J., Huo Z., Cui W., Wang X., Shi J., et al., 2021, The Astrophysical Journal Supplement Series, 257, 65

Szegedi H., Charles P. A., Meintjes P. J., Odendaal A., 2022, Monthly Notices of the Royal Astronomical Society, 513, 4682

Szkody P., 1998, in ASP Conference Series Vol. 137, Spectroscopy of cataclysmic variables: Whopping clues from wiggly lines. p. 18

Szkody P., Dicenzo B., Ho A. Y. Q., Hillenbrand L. A., van Roestel J., Ridder M., DeJesus Lima I., Graham M. L., Bellm E. C., Burdge K., et al., 2020, The Astronomical Journal, 159, 198

Szkody P., Olde Loohuis C., Koplitz B., van Roestel J., Dicenzo B., Ho A. Y. Q., Hillenbrand L. A., Bellm E. C., Dekany R., Drake A. J., et al., 2021, The Astronomical Journal, 162, 94

Sánchez-Sáez P., Reyes I., Valenzuela C., Förster F., Eyheramendy S., Elorrieta F., Bauer F. E., Cabrera-Vives G., Estévez P. A., Catelan M., et al., 2021, The Astronomical Journal, 161, 141

Thorstensen J. R., Motsoaledi M., Woudt P. A., Buckley D. A. H., Warner B., 2020, The Astronomical Journal, 160, 70

Towards Data Science, 2020, A Comprehensive Guide to Convolutional Neural Networks: The ELI5 Way, `https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53`

Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., Altman R. B., 2001, Bioinformatics, 17, 520

Udalski A., Szymański M. K., Szymański G., 2015, Acta Astronomica, 65, 1

Van der Maaten L., Hinton G., 2008, Journal of machine learning research, 9

van Roestel J., Creter L., Kupfer T., Szkody P., Fuller J., Green M. J., Rich R. M., Sepikas J., Burdge K., Caiazzo I., et al., 2021, The Astronomical Journal, 162, 113

van Roestel J., Duev D. A., Mahabal A. A., Coughlin M. W., Mróz P., Burdge K., Drake A., Graham M. J., Hillenbrand L., Bellm E. C., et al., 2021, The Astronomical Journal, 161, 267

van Roestel J., Kupfer T., Green M. J., Wong T. L. S., Bildsten L., Burdge K., Prince T., Marsh T. R., Szkody P., Fremling C., et al., 2022, Monthly Notices of the Royal Astronomical Society, 512, 5440

Verbunt F., 1982, Space Science Reviews, 32, 379

Verbunt F., Zwaan C., 1981, Astronomy and Astrophysics, 100, L7

Vu D. H., Muttaqi K. M., Agalgaonkar A. P., 2015, Applied Energy, 140, 385

Wakamatsu Y., Thorstensen J. R., Kojiguchi N., Isogai K., Kimura M., Ohnishi R., Kato T., Itoh H., Sugiura Y., Sumiya S., et al., 2021, Publications of the Astronomical Society of Japan

Warner B., 1995, Cataclysmic Variable Stars. Cambridge Astrophysics, Cambridge University Press, Cambridge

Wen Q., Sun L., Yang F., Song X., Gao J., Wang X., Xu H., , 2020, Time Series Data Augmentation for Deep Learning: A Survey

Wenger M., Ochsenbein F., Egret D., Dubois P., Bonnarel F., Borde S., Genova F., Jasniewicz G., Laloë S., Lesteven S., et al., 2000, Astronomy and Astrophysics Supplement Series, 143, 9

Wheatley P. J., Mauche C. W., Mattei J. A., 2003, Monthly Notices of the Royal Astronomical Society, 345, 49

Whitehurst R., 1988, Monthly Notices of the Royal Astronomical Society, 232, 35

Wilson R. N., 1991, Contemporary Physics, 32, 157

Woosley S., Bloom J., 2006, Annu. Rev. Astron. Astrophys., 44, 507

Wu K., Kiss L. L., 2008, Astronomy & Astrophysics, 481, 433

Zhang H., 2004, in The Florida AI Research Society The optimality of naive bayes

Zhang Z., 2016, Annals of translational medicine, 4, 218

Özdönmez A., Ege E., Güver T., Ak T., 2018, Monthly Notices of the Royal Astronomical Society, 476, 4162

Šimon V., 2002, Astronomy & Astrophysics, 382, 910

Šimon V., 2021, Monthly Notices of the Royal Astronomical Society, 505, 161