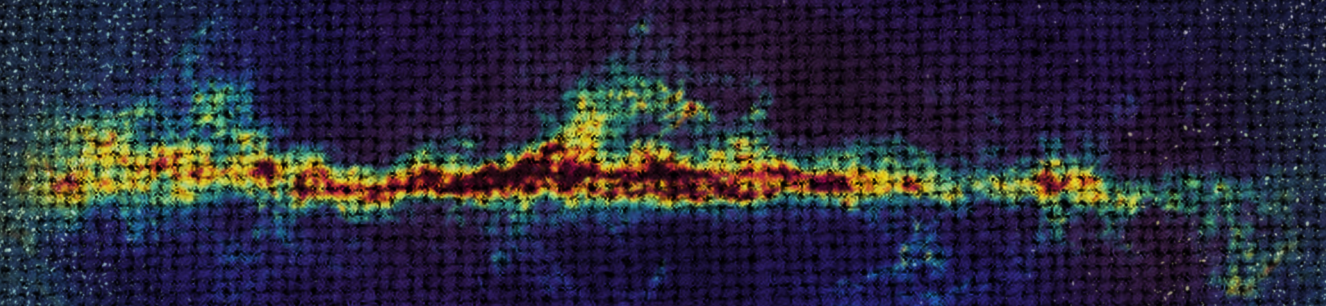
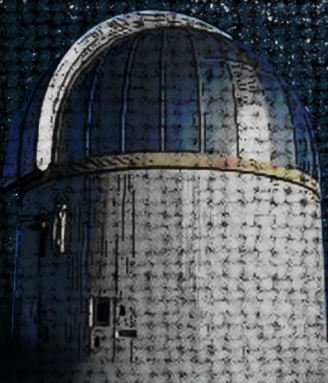


Searching for and Characterising Massive Stars with Machine Learning Techniques



Nicolás Monsalves



Searching for and characterising massive stars with machine learning techniques

Nicolas Matias Monsalves Gonzalez



UNIVERSIDAD DE LA SERENA CHILE

Doctoral Thesis

To fulfill the requirements for the degree of
Doctor of Philosophy in Astronomy
at Universidad de La Serena under the supervision of
Dr. Marcelo Daniel Jaque Arancibia, Universidad de La Serena

Nicolas Monsalves Gonzalez

Acknowledgments

There are a lot of people I want to thank. It has been a little more than eight years in my astronomy career, and I truly couldn't have done it without the support of so many along the way. I will start with academic acknowledgements, and then move on to personal ones.

First, I want to thank my advisor, Dr. Marcelo Jaque (Marce). I met Marce in my third year of undergraduate studies, while I was looking for guidance to begin my final project. I never imagined that our relationship would become so lasting and meaningful. During these years, Marce gave me the freedom to explore whatever I wanted, always supporting me through all my wild ideas. I feel that you have been the wind that helped me navigate as fast as I wished and as far as I needed. I couldn't be more grateful to have met you and to have had you by my side. Over the years, I've experienced many highs and lows, and you have been there in every possible way. I really appreciate that you always reminded me that this journey was mine, that both the undergraduate and PhD paths, as well as the thesis, belonged to me. Of course, it has been challenging to manage that level of freedom, but thanks to it, I now feel like a highly independent researcher. I am sure you were the best person to connect with my personality and to support me in becoming the best version of myself.

Second, but no less important, I want to thank Dr. Amelia Bayo, my unofficial co-supervisor. It is hard to put into words how grateful I am to have met you. You have shown me the kind of scientist I aspire to become: deeply passionate and honest, and that is truly inspiring. Before meeting you in person, I watched an interview in which you said that, for you, science is like a game where nature gives you clues, and you have to piece the puzzle together. I have truly enjoyed playing this game with you. Thank you for always being there, for helping me reach Germany, and for standing by me through every academic and personal challenge. Your unwavering support helped me discover just how vast, beautiful, and full of possibility the world truly is. For me, you are my lighthouse, the one who helped me stay focused and not drift away, and I am sure that none of this would have turned out as well as it has without you.

Third, I want to thank Dr. Rodolfo Angeloni, Dr. Facundo Gómez, and Dr. Guillermo Damke. During my undergraduate studies, Guille and Angeloni were my professors for observational and theoretical astrophysics courses, respectively. From the very beginning, they believed in me, and since then, they have been a constant source of support, whether through recommendation letters for applications to research centers and universities, or simply through thoughtful advice when I needed it most. I really appreciate how approachable and encouraging they've always been. I met Facu at the beginning of my PhD, when he joined my thesis committee. I want to thank you for your essential guidance throughout this journey and for your support in projects beyond my main area of research. I feel very lucky to have had all three of you not only as professors, but later also as collaborators in exciting and rewarding projects. It means a lot to me that our professional paths have continued to cross in such meaningful ways.

Dr. Paula Sánchez, it is impossible not to dedicate a special paragraph to you. I met you online at the La Serena School for Data Science in 2021, where you were my Deep Learning professor. At the time, I asked if you could help me with some problems related to my thesis. I had learned most of the methodologies on my own and barely spoke English. You were about to move to Garching for your fellowship, and now you are staff, which makes me so proud of you. You told me to send you an email in three months, and I did. I was so happy when you replied. I think I even made a story for my close friends on Instagram. From that point, our collaboration began. We later asked if you would join my PhD committee, and since then, you have been an important part of my scientific development. You were also the one who introduced me to Amelia, and thinking about the butterfly effect, it is incredible to realize how much of an impact you have had on my life. We eventually met in person in Germany on New Year's Eve, at your apartment. You were in the kitchen, and I was dancing with your husband (I love Chichi!). I am incredibly grateful to have someone as brilliant and kind as you in my life as a close friend. Thank you for believing in me from the very beginning, for being there during my first experience abroad, and for continuing to support me after I returned. Thank you for always taking care of me.

Contents

Acknowledgements	3
	Page
1 Introduction	7
1.1 Massive Stars: Multi-scale Impacting Objects	7
1.1.1 Main observable: Spectra	11
1.2 Main observable: Variability	14
1.3 Techniques: machine learning to exploit time series data	17
1.4 Scientific goals and outline	20
2 Application of Convolutional Neural Networks to time domain astrophysics. 2D image analysis of OGLE light curves	22
2.1 Introduction	23
2.2 Data	26
2.2.1 Data and pre-procesing	26
2.3 Methodology	28
2.3.1 Data representation	28
2.3.2 Train validation and test sets	31
2.3.3 Convolutional neural network	35
2.3.4 Training process	36
2.4 Results	39
2.4.1 Training performance	39
2.4.2 Metrics	40
2.4.3 Class performance	40
2.4.4 Balancing strategies	41
2.4.5 Convolutional neural network with additional information	42
2.5 Discussion	43
2.5.1 Comparison with previous works	43
2.5.2 Computational resources	48
2.6 Summary and conclusions	50

3	Astrophotometric search for massive stars in the Milky way. Confronting Random Forest predictions with available spectroscopy.	52
3.1	Introduction	53
3.2	The data	55
3.2.1	OBA GDR3: Parental catalog	55
3.2.2	OBA GDR3: Features description	56
3.2.3	OBA GDR3: Spectral type compilation	56
3.3	Methodology: Balanced Random Forest	62
3.3.1	Preprocessing Features	63
3.3.2	Metrics	63
3.3.3	Building the model	64
3.4	Results: Candidates from BRF	65
3.4.1	Performance of the BRF	66
3.4.2	Moderate to high probability massive stars candidates	68
3.5	Discussion	70
3.5.1	Multiplicity considerations	70
3.5.2	Homogeneous validation dataset: LAMOST spectroscopy	70
3.5.3	Comparison with state-of-the-art massive star data bases and catalogs	77
3.6	Summary and Conclusions	78
4	The variability of massive stars	80
4.1	Introduction	81
4.2	Sample, data and adaptation	82
4.2.1	OBA GDR3	82
4.2.2	The TESS data	82
4.2.3	Adaptation tests and searching for periodicity	84
4.2.4	Period extraction	85
4.3	Results	88
4.4	Discussion	89
4.4.1	Methodological Refinements	91
4.5	Summary and Conclusions	92
5	Summary and conclusion	93
6	Appendix	96
A	Effects on the spectral type threshold in the BRF	101

Chapter 1

Introduction

1.1 Massive Stars: Multi-scale Impacting Objects

Massive stars shape the environment and dynamics from their immediate neighbourhoods to galactic distances and even provides us with key knowledge of the distant universe (Eldridge & Stanway, 2022). It is generally accepted that the stellar mass distribution in our Galaxy peaks at $\sim 0.5M_{\odot}$, with steep slopes toward both the high- and low-mass domains (Bastian et al., 2010). Following this framework, in the local universe, less than 1% of stars are expected to be born as massive stars (Kroupa & Weidner, 2003). Despite the lower numbers in our Galaxy, the effect of massive stars in their surroundings is critical. These stars play a fundamental role in astrophysics, even influencing the processes of star and planet formation (Bally & Zinnecker, 2005).

Star formation begins within large, cold giant molecular clouds of $2 - 15$ pc, with temperatures around 10 K and densities of $50 - 500 \text{ cm}^{-3}$ (Bergin & Tafalla, 2007). Stars form through the gravitational collapse of dense cores, which are often embedded in filamentary structures. The formation of massive stars requires denser and more compact clumps of H_2 with column densities between $10^{23} - 10^{24} \text{ cm}^{-2}$ (Zinnecker & Yorke, 2007). Unlike low-mass stars, which tend to form in more isolated environments, massive stars, seem to form in highly clustered and chaotic regions (Beuther et al., 2025). Consequently, high-mass star-forming regions exhibit distinct differences in density, velocity dispersions, and dynamical accretion flows across different spatial scales. It remains uncertain whether massive star formation is a scaled-up version of low-mass star formation or others mechanisms are involved (Rosen et al., 2020).

The gravitational collapse translates into high density cores in which hydrogen fusion is triggered, reaching a stage called the Zero Age Main Sequence (ZAMS). As previously mentioned, the process of star formation is particularly efficient in producing low-mass stars that will spend most of their lives in the main sequence (MS). In general though, the duration of the MS phase is inversely proportional

to the star's mass and luminosity ($t_{MS} \propto M/L$). Following the empirical mass-luminosity relation $L \propto M^\nu$, where ν ranges between 2.7 and 4.0, depending on the star's mass (for massive stars ~ 2.7 , see Eker et al. 2015), the MS lifetime scales as $t_{MS} \propto M^{-(\nu-1)}$. For massive stars, this period is very brief. After 10 Myr, a $20M_\odot$ star has left the MS, a $10M_\odot$ star is nearing its end, and a $5M_\odot$ star remains in its early MS stage. These general guidelines apply however mostly to solar metallicity and low-rotating stars, and can drastically differ for different metallicities or rotation regimes (see Brott et al. 2011 for more details).

During this phase, the process that transforms hydrogen into helium is dominated by proton-proton (P-P) reactions (Bethe & Critchfield, 1938) in lower-mass stars and by the CNO bi-cycle in massive stars (Bethe, 1939). The P-P chain has a temperature threshold of 4×10^6 K, while the CNO bi-cycle requires 15×10^6 K, generating an energy of 6.55 MeV and 6.25 MeV per nucleus, respectively (see table 4.1 in Prialnik 2009). The energy generation in CNO bi-cycle process is highly dependent on temperature roughly approximated by a steep power law ($q_{CNO} \propto \rho T^{16}$), leading to the formation of a convective core and a radiative envelope. Due to convection, the material in this region is uniformly distributed, allowing hydrogen-burning products to be transported to cooler regions (Choudhuri, 2010).

The end of the MS is reached when the hydrogen reservoir in the core is depleted. Hydrogen burning then shifts to a shell surrounding the nucleus, which is primarily composed of helium (Prialnik, 2009). The subsequent evolutionary path depends on the star's mass. At this stage, there is no longer sufficient energy to counteract gravitational forces, leading to the loss of hydrostatic equilibrium. For a massive star, the nucleus begins to collapse because it exceeds the Schönberg-Chandrasekhar limit (Chandrasekhar, 1931) and can no longer support the weight of the overlying envelope. This collapse increases the temperature in the nucleus, while hydrogen burning in the shell adds to the helium mass in the core. As the nucleus contracts, the envelope expands and becomes convective.

Eventually, the temperature reaches 10^8 K, initiating helium nuclear reactions in the core (Tayler, 1994). This phase is extremely short compared to the MS, as the energy generated by the triple-alpha process is only 0.6 MeV per nucleon (?). Once helium is depleted, a similar pattern follows, with heavier elements burning in the core while previously fused elements continue burning in surrounding shells. The energy production decreases with each stage: approximately 0.54 MeV per nucleon for carbon, 0.3 MeV for oxygen, and less than 0.18 MeV for silicon (Prialnik, 2009). This process continues until the core is primarily composed of iron, at which point nuclear reactions become endothermic. With no further energy production to resist gravitational collapse, the core collapses, leading to a supernova explosion (Woosley & Weaver, 1986).

Although the general framework of massive star evolution is well established, several factors significantly influence their evolution, including internal mixing, mass loss through stellar winds, and rotation (Langer, 2012).

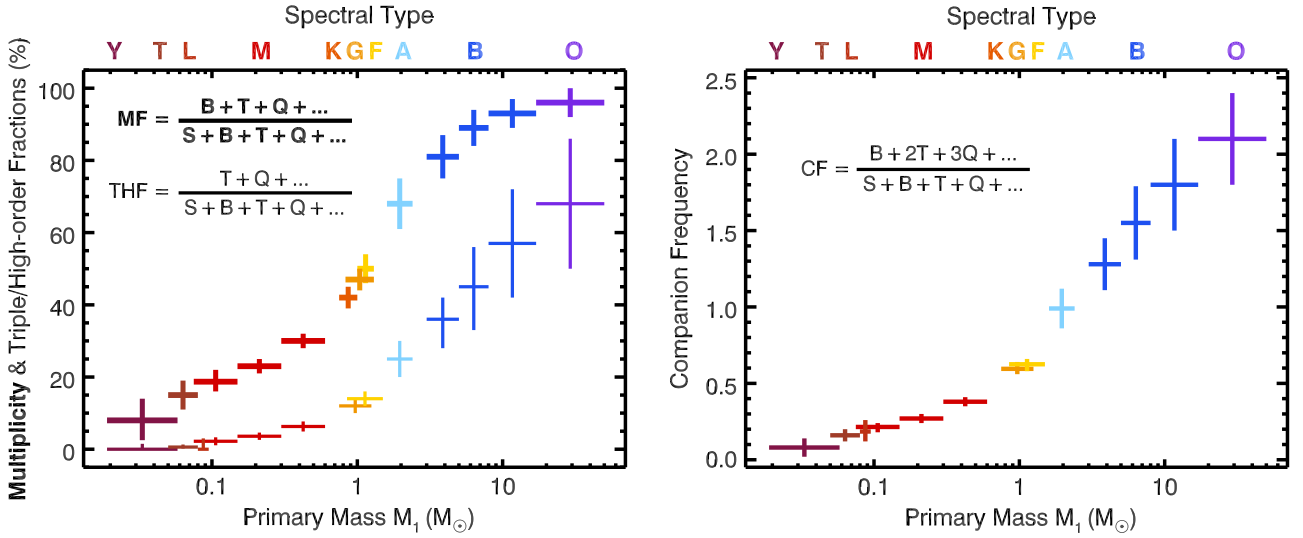


Figure 1.1.1: Bias-corrected multiplicity fraction (MF), triple/high-order fraction (THF), and companion frequency (CF) for Brown Dwarfs and MS stars. Figure extracted from Offner et al. (2023)

In particular, one key assumption of the previously presented simplistic formation and evolution scenario is that of isolation. However, there is plenty of observational evidence invalidating this assumption, as multiplicity has proven to be a fundamental characteristic of massive stars (Sana et al., 2012; Moe & Di Stefano, 2017; Barbá et al., 2017). It is now well established that the majority of massive stars exist in multiple systems, where binary interactions significantly influence their evolution and final outcomes (Sana et al., 2012; Offner et al., 2023). Fig. 1.1.1 shows the bias-corrected values of the multiplicity and triple/high-order fractions for different primary masses. The multiplicity and triple/high-order fractions represent the proportion of primary stars in a sample that have at least one or two companions, respectively. The companion frequency denotes the average number of companions per stellar system within the sample. A monotonic increase in all metrics with stellar mass is observed. In particular, for massive stars ($\sim 8M_\odot$), the multiplicity fraction exceeds 80%, and the companion frequency approaches 1.5, indicating a high prevalence of multiple systems.

The binary nature of massive stars plays a fundamental role in their evolution, with over 70% experiencing mass transfer with a companion (Sana et al., 2012). Among the various interaction mechanisms expected in binary systems, Roche lobe overflow (RLOF) is one of the most significant interactions (Marchant & Bodensteiner, 2024). This occurs when a star expands beyond its Roche lobe, the region where material remains gravitationally bound, leading to mass transfer through the first Lagrangian point. There are still significant uncertainties in this process, particularly concerning how angular momentum and mass are transferred or lost (Packet, 1981; de Mink et al., 2007), and how the companion adjusts its structure to accommodate the incoming mass (Hellings, 1983; Lau et al., 2024). These uncertainties directly impact the properties of post-interaction products, making it more difficult to understand their frequency and characteristics.

When binary interactions are included in the previously presented evolutionary framework, the classi-

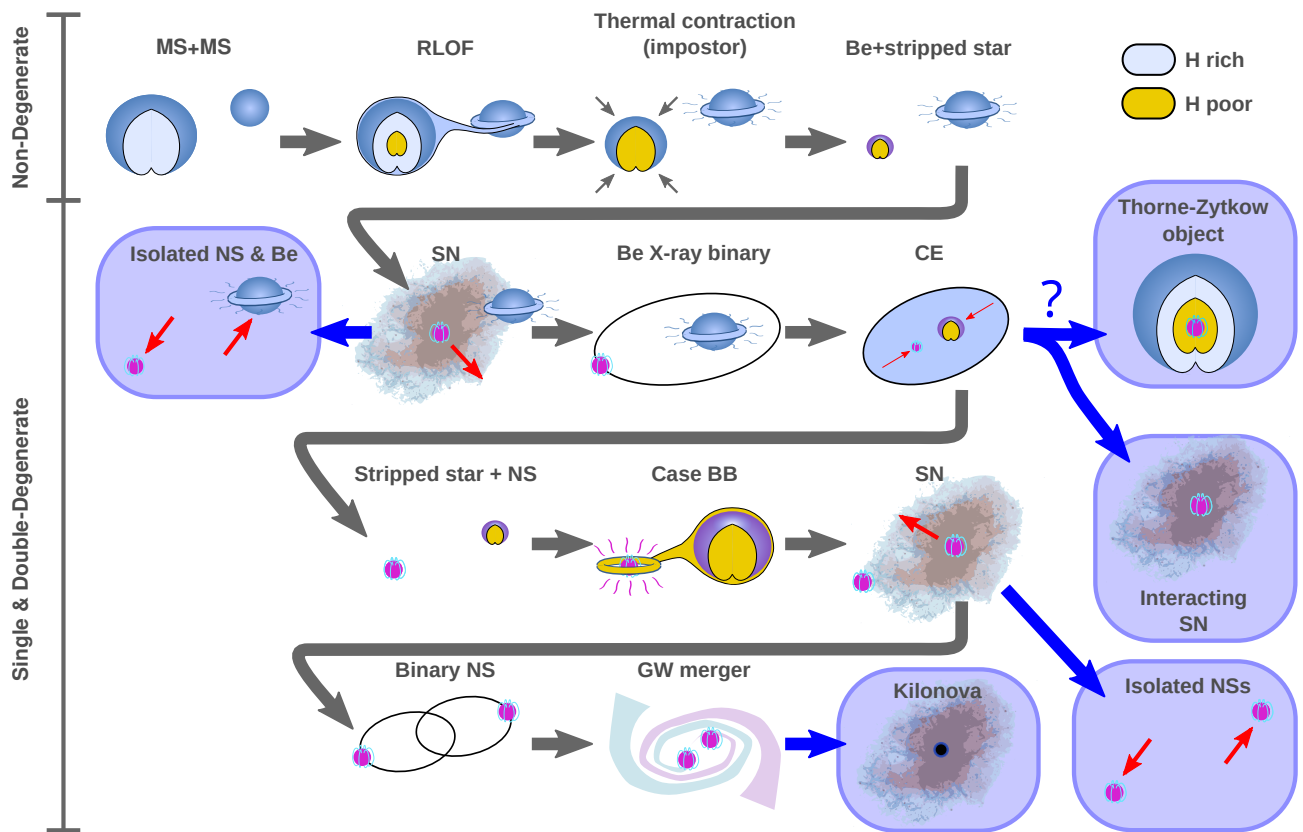


Figure 1.1.2: Depicting the possible evolutionary paths followed by massive binary star. Figure extracted from Marchant & Bodensteiner (2024)

cal path changes significantly. Approximately $30^{+10}_{-15}\%$ of massive main-sequence stars are the product of binary interactions (de Mink et al., 2014). These interactions can lead to various post interaction products, including stripped stars, rapidly rotating accretors, stellar mergers, and double-degenerate binaries (Marchant & Bodensteiner, 2024). Fig. 1.1.2 shows different evolutionary paths for non-degenerate, single-degenerate, and double-degenerate systems. These scenarios originate from two MS stars undergoing RLOF. Post-interaction systems are expected to have unique properties, such as large-scale magnetic fields (Ferrario et al., 2009), the B[e] phenomenon (Miroshnichenko et al., 2007), and unusual rotation rates (Wang et al., 2020b). Although some observational insights have been proposed to identify post-interaction products (de Mink et al., 2014), no definitive observational criteria have been established. Probably the most representative example of a pre-interaction product is a pre-interaction binary (de Mink et al., 2014), which can be identified through large radial velocity variations.

In addition to individual evolutionary differences, multiplicity in the high-mass end of the spectrum also has to be taken into account when conducting integrated population studies. For instance, if the multiplicity is not properly accounted for in population synthesis modeling, the estimation of key parameters such as the star formation rate and the initial mass function (Eldridge & Stanway, 2022) will be heavily affected. For instance, stripped stars can heavily influence the ionizing spectrum, producing the appearance of a younger stellar population (Götberg et al., 2019). In galaxies, massive main-sequence populations include stellar mergers and mass-transfer products, which are often observationally indistinguishable from single stars (de Mink et al., 2014; Henneco et al., 2024; Marchant & Bodensteiner, 2024). Differentiating between these evolutionary outcomes remains a central observational challenge.

1.1.1 Main observable: Spectra

As previously discussed, mass is one key parameter that determines the evolution of stars. Unfortunately, a star's mass does not correlate neither directly nor univocally with any single observable. However, studying spectral characteristics can help bridge the gap between physical parameters and observables.

The general framework for classifying stellar spectra is the MK system (Morgan & Keenan, 1973). This follows an Aristotelian inductive approach, meaning that the classification is based on a representative of the class that define the system. In the case of the MK system, these classes correspond to standard stars. These stars serve as the foundation of the system, and the classification scheme involves comparing the stellar spectra of new stars with those of the standards.

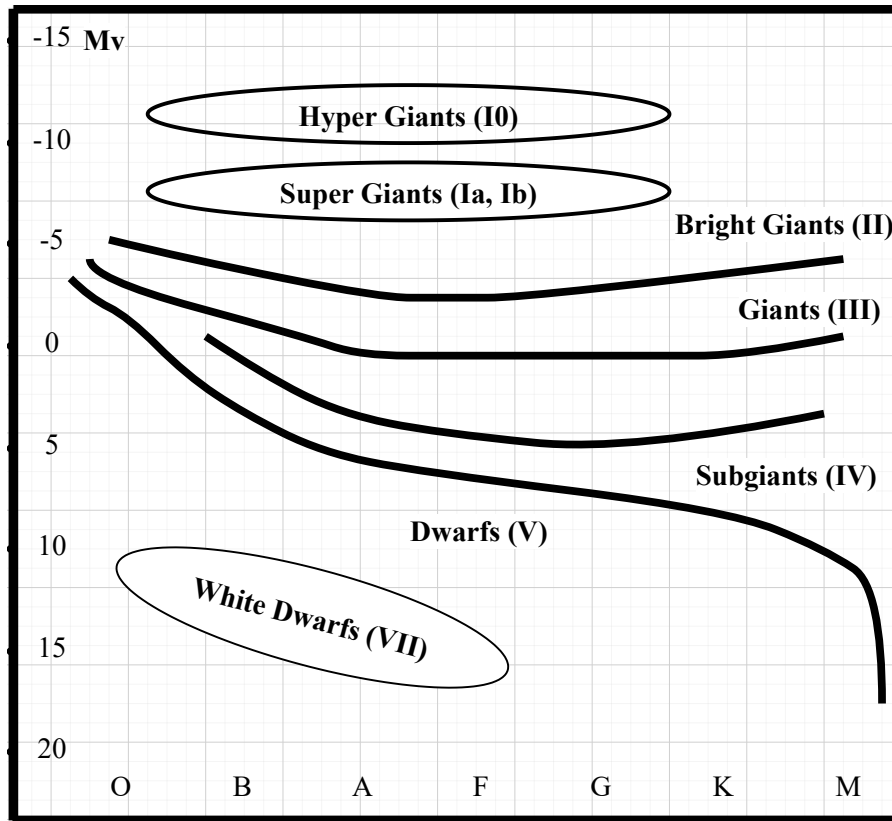


Figure 1.1.3: Schematic representation of the Hertzsprung–Russell diagram showing absolute visual magnitude versus spectral type, with luminosity classes labeled using Roman numerals.

Being a purely empirical approach, matching the stellar features to representatives of discrete classes, this system is independent of theoretical models but, as a consequence, is also not directly linked to physical parameters. However, the sequence of standards, and hence labels, shows a strong correlation with effective temperature. Over time, the system has evolved in two ways: by refining these categories with additional stars within existing classifications and by extending the classification beyond its original boundaries (Maíz Apellániz et al., 2024).

The MK system consists of a spectral type sequence of seven letters, ordered by decreasing temperature: O, B, A, F, G, K, and M, following the Harvard temperature classification (Cannon & Pickering, 1901). Additionally, numbers from 0 to 9 are assigned within each spectral type, where 0 represents hotter stars and 9 corresponds to cooler ones. In recent decades, the spectral classification system has been extended beyond the classical sequence to include later-type objects, such as L, T, and Y dwarfs. Similarly, intermediate subtypes such as 9.5 have also been introduced to allow for finer classifications (Maíz Apellániz et al., 2024).

The system also includes information on the luminosity class: I for supergiants, II for bright giants, III for giants, IV for subgiants, and V for dwarfs. Finally, additional symbols or notations can be included to describe peculiarities in the spectrum. Fig. 1.1.3 shows a variation of the HR diagram

using spectral type and luminosity class information.

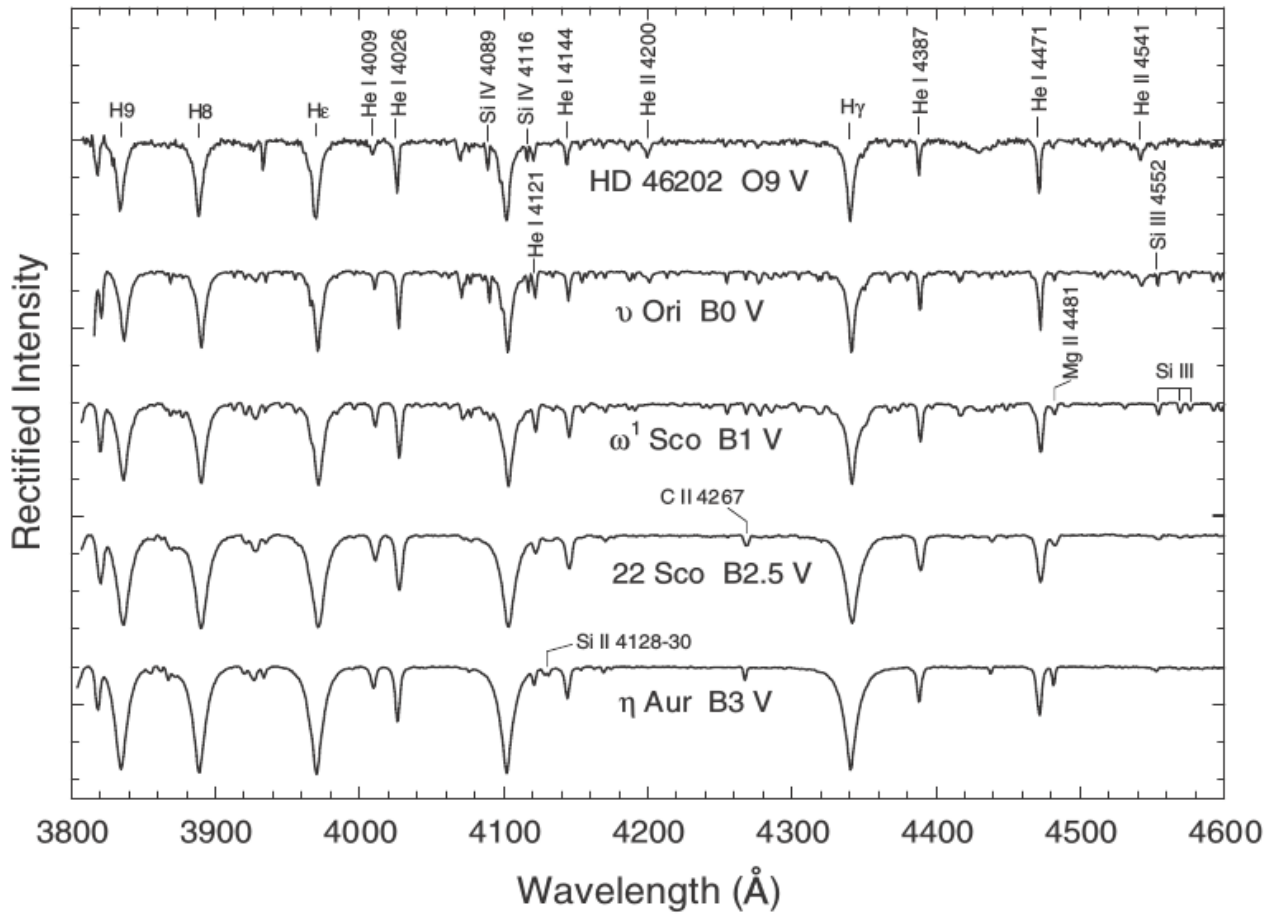


Figure 1.1.4: Stellar spectra for massive dwarf stars, arranged from top to bottom in order of decreasing temperature, from O9 star to B3. Figure extracted from Gray & Corbally (2009)

Massive stars cover spectral types from O to B2 for dwarfs. For evolved stars, this range extends to O–B5 for giants and O–B9 for supergiants (Morgan, 1951), as these stars are expected to evolve from hotter dwarfs. The dominant spectral features include the Balmer hydrogen series, helium lines, and metallic lines from elements such as carbon (C), nitrogen (N), oxygen (O), and silicon (Si). Fig. 1.1.4 shows the spectra of dwarf stars from O9 to B3 in the blue part of the optical spectrum, that is the region historically used for spectral classification given the high-sensitivity of photographic plates in this wavelength range (Pickering, 1890; Walborn, 1972; Gray & Corbally, 2009; Sota et al., 2011, 2014).

Expanding on the previous description, for O and B stars, the Balmer lines are among the most easily identifiable features, dominating the stellar spectrum. In addition, He II lines are especially strong in O-type stars and, together with He I lines, are used to determine subspectral types. The ratio of He II $\lambda 4541$ to He I $\lambda 4471$ serves as a primary classification criterion, being approximately unity at spectral type O7 and appearing stronger or weaker for earlier or later types, respectively (Gray & Corbally, 2009). Additional criteria are used for earlier and later type in the sequence (Maíz Apellániz

et al., 2024). Luminosity class is based on the Of phenomenon in which the He II λ 4686 evolves from strong absorption in dwarfs to emission in supergiants (Walborn, 1973). Additionally, N III emission lines increase in intensity with luminosity. For a detailed study on the spectral classification of O-type stars, we refer to Sota et al. (2011).

For B-type stars, the mapping between spectral typing and physical parameters suffers from a higher degree of degeneracy. Early B-type stars are massive across both luminosity classes, whereas late B-type dwarfs are not. Consequently, objects of very different nature can share the same spectral type. The primary spectral indicators used for classification include He I lines, H lines, and Si lines (Maíz Apellániz et al., 2024). The He I/ Mg II line ratio is useful for helium-normal star later than B3 (Gray & Corbally, 2009), being unity close B6 - B7 (Ramírez-Preciado et al., 2020). The most commonly used luminosity criterion involves the Si/He ratio, which shows a positive correlation with luminosity. However, careful consideration must be given to spectral peculiarities (see (Maíz Apellániz et al., 2024) for a recent review). Additionally, the H lines, in terms of both width and equivalent width (EW), increase with spectral subtype and luminosity class (Negueruela et al., 2024).

1.2 Main observable: Variability

Stellar variability is a ubiquitous phenomenon in the life of stars, in which the light we record from them changes over time. The variations range in amplitude from millimagnitude (mostly reachable with space based observations) to several magnitudes (like the case of strong transients easily detected from ground-based observations). The timescales of variability are also very broad, spanning from milliseconds to years. The observed changes in brightness are stored in a light curve (LC), which traces the flux variation as a function of time. The study of different types of variability provides insightful information about the processes responsible for these changes, which can be related to the nature or evolutionary status of the star (Percy, 2007).

Figure 1.2.1 shows a phenomenological variability tree that seeks to organize different types of variability. The first level distinguishes between extrinsic and intrinsic variability, depending on whether the phenomenon is caused by physical changes in the object or by geometrical effects that alter our perception of its brightness. The second level, less phenomenological, roughly groups (non-resolved) objects by their nature (Asteroids, Stars AGN). The third level describes the mechanism responsible for the variation. Further levels are present to different depths depending on the uniqueness of the characteristics of the given mode of variability.

Deeply exploring all kinds of variability presented in Fig. 1.2.1 is beyond the scope of this chapter and we will only focus on short descriptions of the types related to massive stars.

Intrinsic variability includes cataclysmic events, eruptive events, and pulsations. Cataclysmic events can be either an interacting binary system with a white dwarf, or a star exhibiting large-amplitude

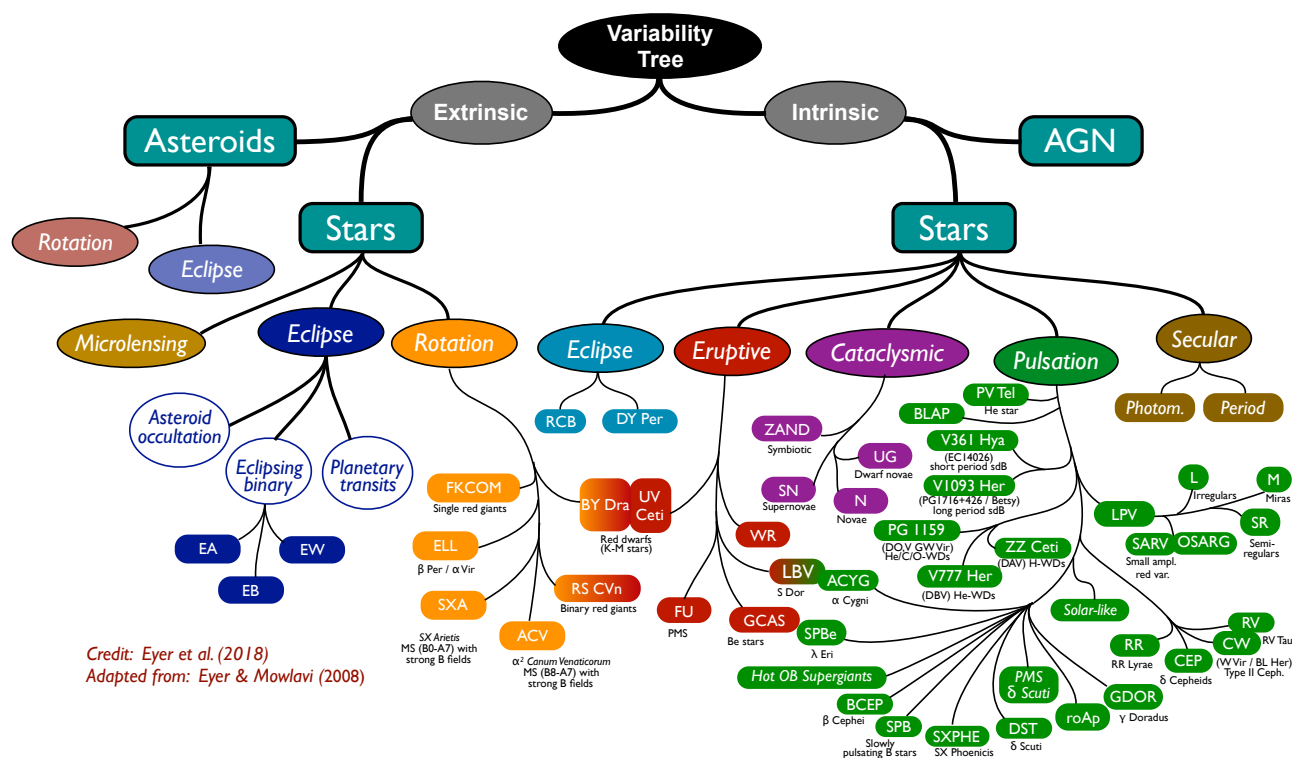


Figure 1.2.1: Variability tree diagram showing the different causes of stellar variability. Figure extracted from Gaia Collaboration et al. (2019).

outbursts. One example related to the final evolution of massive stars are the Type II supernovae explosion (Smartt, 2009).

The eruptive events can be related to mass ejection or chromospheric activity. Among massive stars, examples include classical Be stars, that are officially called: γ Cass stars. They are rapidly rotating, non radially pulsating B type stars with a circumstellar decretion disk (Rivinius et al., 2013). Wolf-Rayet stars (WR), characterized by broad and strong emission lines (Wolf & Rayet, 1867), exhibit irregular light variations caused by unstable mass outflow from their atmospheres. Luminous Blue Variables (LBVs) are evolved massive stars that display distinctive variability or eruptive behavior that sets them apart from other stellar types (Weis & Bomans, 2020).

Pulsation variability consists of mechanical oscillations of the stellar interior that respond to perturbations of its equilibrium state. These classes include numerous subdivisions that can be defined by: radial and non-radial pulsators, the excitation mechanism that triggered the oscillation, evolutionary status, and mass (Catelan & Smith, 2015). For massive stars' pulsation, in Fig. 1.2.1, we can expect massive stars to populate the classes: β Cephei stars, which are young Population I, near-

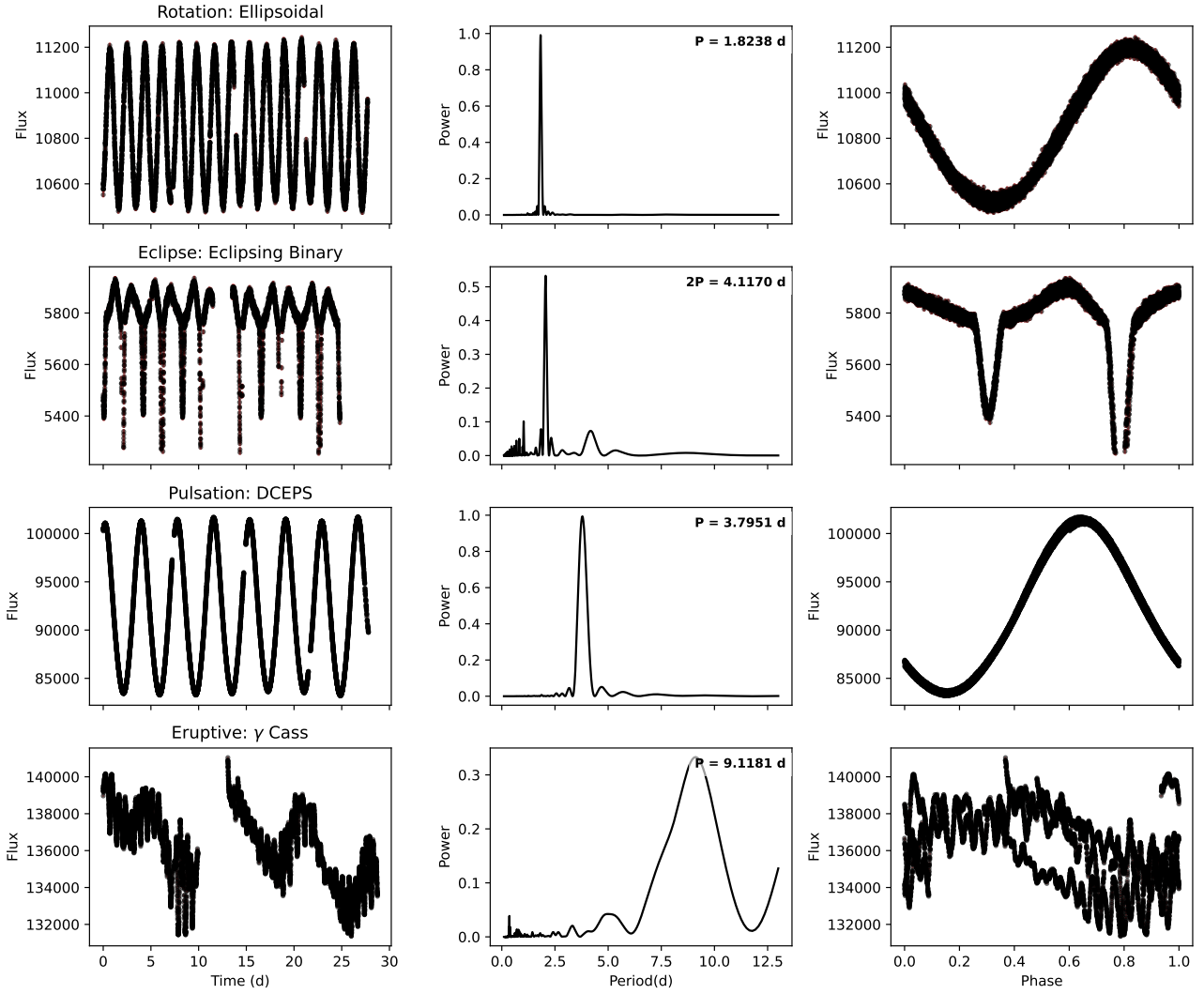


Figure 1.2.2: Different LCs of massive stars from TESS. The second column displays the Lomb-Scargle periodogram. The third columns shows the phasefolded LC (see equation 2.3.1)

main-sequence stars with masses between $8\text{--}18 M_{\odot}$ (Aerts et al., 2010); Slowly Pulsating B stars (Waelkens, 1991); and pulsations observed in hot OB supergiants (Bowman et al., 2019). Since the analysis of pulsation in stars opens a unique window to the study of interiors, the availability of data from Transiting Exoplanet Survey Satellite (TESS) (Ricker et al., 2015) is meant to revolutionise our understanding in this area (Bowman, 2023).

The extrinsic variability of massive stars is strongly related to the presence of one or more companions, due to their high degree of multiplicity, as in eclipsing and ellipsoidal binaries. Ellipsoidal binaries are non-eclipsing close binaries in which the components are distorted by their mutual gravitation (Morris, 1985). Changes in the LC are observed due to surface deformation of the stars, typically producing a near-sinusoidal modulation.

Eclipsing binaries are found in systems where at least two stars are gravitationally bound in a close

orbit (typically $a \lesssim 0.5$ AU (Moe & Di Stefano, 2017)), with the orbital plane aligned with the line of sight. In such configurations, the observer can detect periodic decreases in brightness due to mutual occultation. In addition, spectroscopic variability can be observed due to Doppler shifts in the spectral lines. These systems are classified as single-lined (SB1) or double-lined (SB2) spectroscopic binaries, depending on whether the spectral lines of both components can be resolved. Eclipsing binaries are especially valuable because their combination with double-lined spectroscopic data allows for the determination of fundamental, model independent astrophysical parameters (Torres et al., 2010; Barbá et al., 2022; Tkachenko et al., 2024; Martín-Ravelo et al., 2024). Orbital parameters such as primary mass, mass ratio, orbital period, and eccentricity provide critical insights into various evolutionary stages of stars (Moe & Di Stefano, 2017; Offner et al., 2023). However, the total population of fully characterized systems remains a major bottleneck in the field (Eker et al., 2014).

Figure 1.2.2 shows LCs exemplifying of different types of variability in a sample of massive stars. The LCs are extracted from TESS. From top to bottom, we show examples of ellipsoidal binaries, eclipsing binaries, Delta Cepheids (DCEP), and γ Cass stars. The right panel displays the variability over approximately 30 days, which corresponds to the baseline of one sector in the TESS mission. The second panel shows the Lomb-Scargle periodogram of the time series.

1.3 Techniques: machine learning to exploit time series data

Over the last decade, a vast amount of data have become available in astronomy, spanning a broad range of the electromagnetic spectrum and even extending beyond it (multi-messenger), originating from both ground-based and space-based telescopes. As a natural response, machine learning (ML) approaches have become increasingly popular, as they offer efficient, scalable, and accurate solutions to the problem of massive, complex and diverse data analysis.

ML is a subfield of artificial intelligence that aims to develop computer programs which automatically improve their performance in specific tasks with experience (Mitchell & Mitchell, 1997). Conventional ML techniques require human-lead domain-specific feature extraction to convert natural data into numerical representations suitable for learning algorithms. The rise of powerful computational resources, such as Graphics Processing Units (GPUs) that support the training of models with higher complexity, has enabled deep learning to emerge as a key subset of ML (Zhang et al., 2018). It allows models to be trained directly on natural data, automatically learning/extracting relevant features without the need for manual design by human engineers.

We can broadly distinguish two main branches of Machine Learning (ML): supervised learning and unsupervised learning. In supervised learning, the algorithm (deep or classic) learns from examples, which requires the availability of ground truth information, commonly referred to as labeled data. Once the relation(s) between the data and the classes is learned, this knowledge can be used to infer predicted labels in continuous or discrete spaces. In contrast, unsupervised learning does not rely on

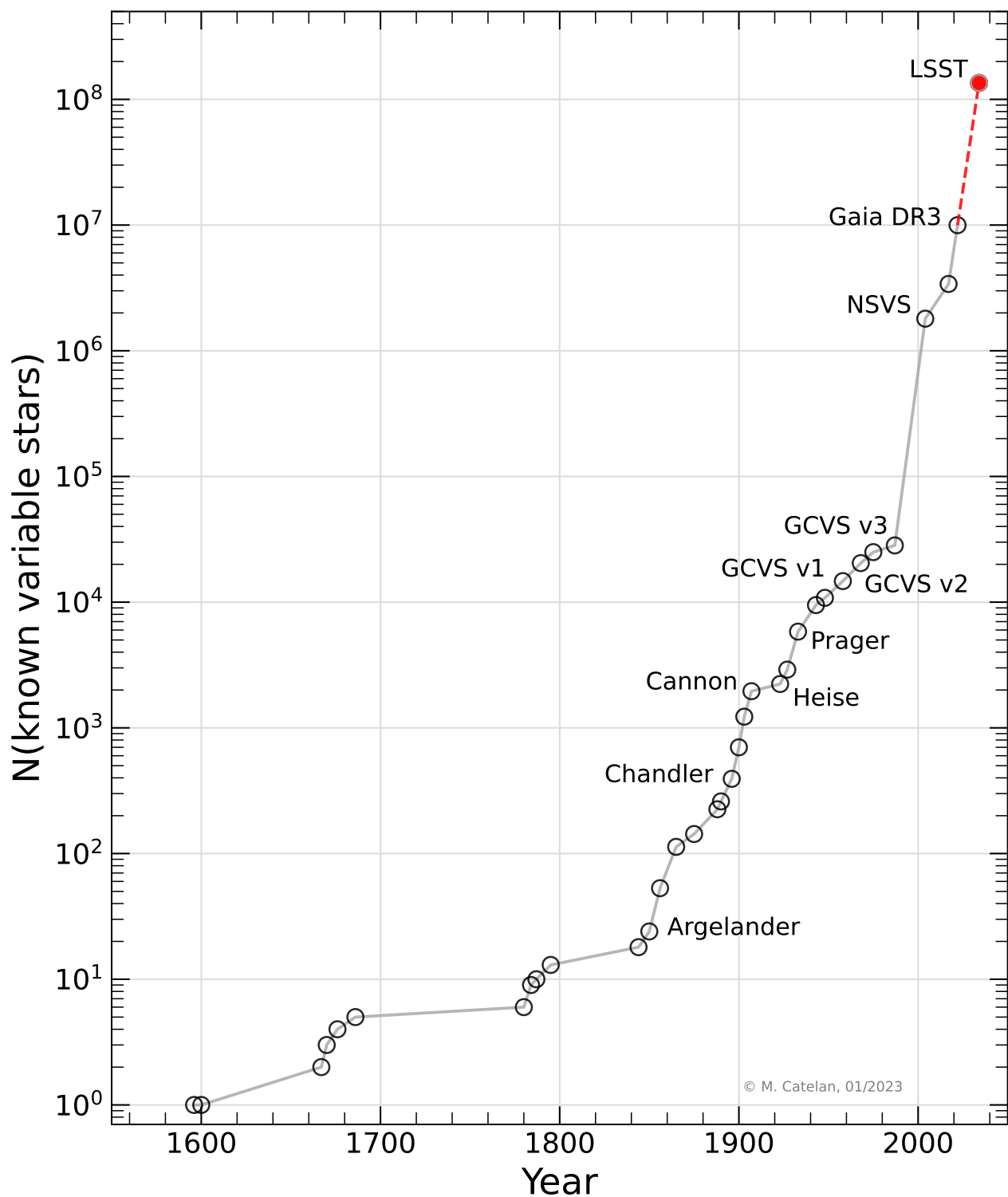


Figure 1.2.3: Number of variable stars per year. Figure extracted from Catelan (2023)

a representative set of labeled examples. Instead, the goal is to identify patterns or structures within the data itself.

The training process of any model requires a critical division of the available data (labeled in the case

of supervised learning) into subsets with clear purposes. In the case of classical ML, the training data is (disjointly) separated into training and testing sets. In the case of deep learning, the separation (also disjoint) is threefold: training, validation and testing. The training set is used to fit the hyperparameters of the model that best reproduce either the classes of those data (supervised) or the underlying structures (unsupervised). The performance of this model is then evaluated independently on the test data. In the case of deep learning, it is necessary to have an indirect evaluation during the training process, which is provided by the validation set. Empirically, we can separate the sets into 80% for training and 20% for testing, or 70% training, 15% validation, and 15% testing; However, the exact split depends on the scientific case and the amount of available data. If the test set is sufficiently representative, more data can be allocated to training.

The application of ML to astrophysical research has been ongoing for more than 30 years. Some early efforts focused on classifying different morphology of galaxies (Storrie-Lombardi et al., 1992) and galaxy spectra (Folkes et al., 1996), as well as estimating physical parameters from stellar spectra (Bailer-Jones et al., 1997). At present, similar challenges are being addressed using larger datasets and more sophisticated machine learning techniques.

Regarding time-domain astrophysics, there has been an explosion of available data due to wide-field surveys (Catelan, 2023), a trend that will only intensify with the start of operations of the *Vera C. Rubin* Observatory and its Legacy Survey of Space and Time (Ivezić et al., 2019). Fig. 1.2.3 shows the number of variable stars per year, with a significant increase around the year 2000 and an expected order-of-magnitude rise from the current number of known variable stars with LSST.

To characterize, analyze, and classify this large volume of data, various efforts have been made available. For example, a widely used approach for classifying variable stars involves first extracting features that summarize the information content of the LCs. These features are then used as input to the algorithm within a supervised learning framework (Belokurov et al., 2003; Woźniak et al., 2004b; Debosscher et al., 2007; Blomme et al., 2010; Richards et al., 2011; Dubath et al., 2011; Farrell et al., 2015; Bass & Borne, 2016; Sánchez-Sáez et al., 2021). Unsupervised methods have also been explored for the characterization of variable stars (Brett et al., 2004; Matijević et al., 2012; Valenzuela & Pichara, 2018).

One advantage from the feature extraction methodology, is that the man-designed features try to recover our pre-defined ideas of the variability characteristics that we try to interpret physically. This process, however, is often time-consuming and requires substantial integration effort to generalize across different surveys (Aguirre et al., 2019). An alternative approach involves deep learning methods. Through the use of supervised learning techniques (Aguirre et al., 2019; Becker et al., 2020; Jamal & Bloom, 2020; Szklenár et al., 2022; Olmschenk et al., 2024; Monsalves et al., 2024) and Unsupervised techniques (Pantoja et al., 2022; Chan et al., 2022; Ding et al., 2024). However, the interpretation of the results of these methodologies is more complex than the man-designed features.

Within the supervised learning framework, both traditional and deep learning methods exhibit a strong dependence on the quality and characteristics of the training data. Consequently, any bias present in the training set will very likely be reflected in the model, potentially affecting its predictions. Therefore, careful consideration must be given to the astrophysical interpretation of the results and the criteria used for sample selection (Richards et al., 2012).

A key challenge of supervised learning is class imbalance. When a training set is highly imbalanced and no mitigation strategy is applied, the majority class will dominate the learning process, and the algorithm will perform poorly on the minority class (Chawla et al., 2004). This is particularly relevant in astronomy, where the distribution of astrophysical phenomena is inherently non-uniform, either due to intrinsic rarity or observational limitations that bias detection toward specific types. In the following chapters, different strategies to address class imbalance will be presented.

1.4 Scientific goals and outline

The overall aim of this thesis is to improve the astrophysical knowledge regarding massive stars as a population. This is accomplished by delivering a large sample of candidate and spectroscopically confirmed massive stars with characterized stellar variability. This work focuses on automation and optimization of workflows to identify and characterize new massive stars from existing surveys, in preparation for the forthcoming influx of time-series and spectroscopic data from next-generation surveys. All results are developed with machine learning methodologies or automated methods, aiming to evaluate and refine the benefits and limitations in the massive stars field, traditionally defined by the manual analysis of highly complex astrophysical phenomenon.

In Chapter 2, we introduce a new deep learning methodology based on a convolutional neural network to classify different types of periodic variable stars. We cover a broad range of variability across different masses and evolutionary states. This framework relies on image-based recognition of variability, mimicking the traditional approach of visually inspecting LCs. We then combine the image features with period and amplitude information to enhance the classification results. The method is highly optimized for both speed and accuracy, enabling the efficient processing of millions of LCs within minutes.

In Chapter 3, we present a catalog of candidate and spectroscopically confirmed bright massive stars in the Galaxy. We performed a literature search to identify known massive stars and used this sample in a supervised machine learning approach, incorporating astrometric and photometric information to isolate high-probability candidates. The candidates are externally validated using spectroscopic data, and we describe an automated methodology for line detection and equivalent width calculation. A decision path is developed to classify the sample based on the detection of spectral lines characteristic of standard dwarf stars. The spectroscopic validations are consistent with the expected performance of the method, and the candidate sample is estimated to increase the known population of bright

massive stars by 50%.

In Chapter 4, we present the preliminary results on the characterization of massive stars variability. We used the candidate and spectroscopically confirmed bright massive stars in the Galaxy presented in Chapter 3 to search for high precision space-based photometry. Subsequently, we adapt and apply the convolutional neural network methodology described in Chapter 2 to classify different types of variability. We present the distribution of the various variability types within the sample and discuss the implications of these findings, as well as future work that will be pursued with this dataset.

Chapter 2

Application of Convolutional Neural Networks to time domain astrophysics. 2D image analysis of OGLE light curves

This chapter was originally published as:

Application of Convolutional Neural Networks to time domain astrophysics. 2D image analysis of OGLE light curves.

N. Monsalves, M. Jaque Arancibia, A. Bayo, P. Sánchez-Sáez, R. Angeloni, G. Damke, J. Segura Van de Perre. *Astronomy & Astrophysics*, 2024, 691, A106

Abstract

In recent years the amount of publicly available astronomical data has increased exponentially, with a remarkable example being large-scale multiepoch photometric surveys. This wealth of data poses challenges to the classical methodologies commonly employed in the study of variable objects. As a response, deep learning techniques are increasingly being explored to effectively classify, analyze, and interpret these large datasets. In this paper we use two-dimensional histograms to represent Optical Gravitational Lensing Experiment (OGLE) phasefolded light curves as images. We use a Convolutional Neural Network (CNN) to classify variable objects within eight different categories (from now on labels): Classical Cepheid (CEP), RR Lyrae (RR), Long Period Variable (LPV), Miras (M), Ellipsoidal Binary (ELL), Delta Scuti (DST), Eclipsing Binary (E), and spurious class with Incorrect Periods (Rndm). We set up different training sets to train the same CNN architecture in order to characterize the impact of the training. The training sets were built from the same source of labels but different filters and balancing techniques were applied. Namely: Undersampling (U), Data Augmentation (DA), and Batch Balancing (BB). The best performance was achieved with the

BB approach and a training sample size of ~ 370000 stars. Regarding computational performance, the image representation production rate is of ~ 76 images per core per second, and the time to predict is $\sim 60 \mu\text{s}$ per star. The accuracy of the classification improves from $\sim 92\%$, when based only on the CNN, to $\sim 98\%$ when the results of the CNN are combined with the period and amplitude features in a two step approach. This methodology achieves comparable results with previous studies but with two main advantages: the identification of miscalculated periods and the improvement in computational time cost.

2.1 Introduction

By definition, variable stars exhibit detectable changes in brightness over time. These changes can result from physical processes intrinsic to the stars or geometric processes that affect them. Geometric processes include eclipses due to a companion (Charbonneau et al., 2000) or rotation (Berdyugina, 2005). The intrinsic processes include pulsation, flares (Shibayama et al., 2013), and cataclysmic eruptions (Hellier, 2001).

Time domain astronomy has played a fundamental role in our current knowledge of the Universe (Hubble, 1929; Riess et al., 2016, 2019). Historically, these objects have been a great source of information because they allow us to determine fundamental astrophysical relationships from the nature of variation. One excellent example is the period-luminosity relation discovered by Leavitt & Pickering (1912), which was fundamental to calculate the distance to M31 (Hubble, 1925) thus marking the beginning of extragalactic astronomy (Schneider, 2006).

Based on their lightcurve (LC) morphology, variable stars can be broadly classified into regular, semiregular and irregular variables. Regular variables exhibit a clear pattern that repeats over time, whereas irregular variables show no obvious signs of periodicity. Semiregular variables display some signs of periodicity, along with stochastic variations (Percy, 2007). Aspects of the morphology of different types of LCs can be quantified in so-called features that vary in complexity regarding their computation, from simple amplitudes to more involved accounting of the timescales of variations. Such features encode hints on the structure and evolutionary state of these variable stars (Catelan & Smith, 2015).

In this context, regular variables are particularly useful since the characteristics of their LCs allow us to infer fundamental astrophysical parameters. The period-luminosity relations of pulsating stars, including Cepheids (CEP), RR Lyrae (RR), and Mira (M), play a crucial role in determining the cosmic distance ladder across the universe (Riess et al., 2011; Muraveva et al., 2018; Sanders, 2023). Eclipsing Binary (E), on the other hand, allow us to directly estimate mass, radius, temperature, and absolute luminosity of the system's components. These parameters are necessary to test the models of stellar structure and evolution (Torres et al., 2010). They are the primary source of empirical information about the properties of stars, making them a foundational pillar of modern astrophysics

(Southworth, 2012).

In the last decades, multiepoch photometric surveys have been developed to address different scientific problems. These surveys obtain time series of different objects with an extended temporal coverage, such as: Massive Compact Halo Objects (MACHO) (Alcock et al., 1997), The All Sky Automated Survey (ASAS) (Pojmanski, 2002), The Optical Gravitational Lensing Experiment (OGLE) (Udalski, 2003), The Northern Sky Variability Survey (NSVS) (Woźniak et al., 2004a), The Catalina Real-Time Transient Survey (CRTS) (Drake et al., 2014), The Zwicky Transient Facility (ZTF) (Bellm et al., 2019) and the *Gaia* mission (Gaia Collaboration et al., 2023), just to cite a few. These surveys have generated (or continue to do so) large amounts of data, pushing towards multidisciplinary efforts to develop new methodologies to accurately and efficiently analyze them. This scenario of big data will only increase with the imminent arrival of new telescopes such as the *Vera C. Rubin* Observatory and its Legacy Survey of Space and Time (Ivezić et al., 2019), which will revolutionize the way astronomy works, generating tens of terabytes of data each night (Ivezić et al., 2019).

A myriad of automatic classification methodologies have surged in the community to tackle this challenge. Examples of such methodologies can include conventional supervised machine learning techniques operating over a space of features that characterize the LCs, and subsequently apply these features to the classification of the LC (Pichara et al., 2012; Nun et al., 2015; Sánchez-Sáez et al., 2021). These methodologies have shown robust results. However, feature extraction is a complex process that typically requires significant time and research to be conducted effectively (Aguirre et al., 2019; Sánchez-Sáez et al., 2021).

In the last few years deep learning methods have achieved highly accurate results (Naul et al., 2018; Becker et al., 2020; Martínez-Palomera et al., 2022). These methods automatically extract important features through a trial and error training process. They require large amounts of data for training and arise as a natural response to the current situation of big data. The great popularity of these methods in recent years is due to the current availability of more powerful computational resources, such as Graphics Processing Unit (GPU) (Zhang et al., 2018).

The most common representation of variable stars includes sequential data, statistical / characteristics abstraction of the former, or a combination of both. Sequential data consist of time series or different Adjustments of the time series. For example in Becker et al. (2020) they used the difference with previous measurements in time and magnitude. The tabular data can consist of features describing the LC, either from statistical information or astrophysical knowledge (Nun et al., 2015). Another less explored way of representing LCs is by using an image. Mahabal et al. (2017) use the raw LC with a 2D histogram of the differences in days (dt) versus the differences in magnitude (dm). They proposed an image-based classification of variable stars data taken from the Catalina Real-Time Transient Survey. They obtain performances comparable to Random Forest (RF; Breiman (2001)) without feature extraction, and highlight potential future applications with this approach. More recently, Szklenár

Table 2.1.1: Number of variable stars used in this study from OGLE III and IV time series.

Variability class	Acronym	Non-unique LC from OGLE catalogs	I filter data missing	$n_{obs} < 60$	Selected from OGLE IV	Possible blended star	Final numbers
Ellipsoidal binary	ELL	26880	925	29	0	510	25416
Mira	M	74542	1214	4851	5161	36	63280
Classical Cepheids	CEP	19680	116	481	7638	79	11366
Delta Scuti	DST	45413	532	483	47	4344	40007
Eclipsing binaries	E	516143	14273	699	29381	866	470924
Long period variable	LPV	335255	0	647	122	1055	333431
RR Lyrae	RR	170408	1466	3332	41421	300	123889
Total	-	1188321	18526	10522	83770	7190	1068313

Note. The Col. 1 presents the class of variability. The Col. 3 displays the number of variable stars obtained from the OGLE catalogs. The Col. 4 lists stars without time series in the I-band filter. The Col. 5 indicates stars with fewer than 60 observations. The Col. 6 shows stars that are observed in both OGLE III and OGLE IV. The Col. 7 shows stars eliminated due to an internal match within 1 arcsec. The Col. 8 presents the final count of independent stars. Table .1 shows the OGLE reference for each variability class.

et al. (2020, 2022) studied the classification of phasefolded LCs in OGLE data. They represented the LCs as 8-bit images with a size of 128×128 pixels, using a black background with white plotted dots. The main idea is to simulate the traditional approach to displaying and analyzing time-series data. In their first paper, they presented the methodology for image-based classification, and in their second paper, they extended their work by using a multi-input neural network that combined images with tabular information.

This work is the first in a series focusing on the highly accurate classification of variable stars in big data astronomy. The aim of this initial paper is to further explore image classification. We present the methodology and demonstrate that this approach is suitable in terms of speed, accuracy, and minimizing computing resources. The second paper will concentrate on the adaptability of the model to various surveys, emphasizing the optimization of the number of classified stars necessary for retraining the model in each survey. The paper is organized as follows: in Sect. 4.2 we summarize the OGLE data from different survey missions and explain the download process. We also describe the preprocessing steps used to generate training, validation, and test sets. In Sect. 2.3 we present our Convolutional Neural Network (CNN) architecture and detail the training process. In Sect. 2.4 we present the main results for the classification of variable stars with different balancing techniques. We also propose a combination of tabular and image information to improve our CNN. In Sect. 2.5 we compare our algorithm with others and discuss the computational resources and time required. Finally, in Sect. 4.5, we summarize our work and present future projections on this topic.

2.2 Data

2.2.1 Data and pre-processing

For the development of this work, we utilized the time series data from OGLE. The main objective of OGLE is the search for dark matter with microlensing phenomena (Udalski et al., 1992). However, due to the extensive number of observations, and the prolonged timeframe that can be archived over half a century, there has also been a focus on the identification and classification of variable stars. OGLE has monitored the Large Magellanic Cloud, the Small Magellanic Cloud, the Milky Way disk, and the Milky Way bulge.

The observations were conducted with the 1.3-m Warsaw telescope at Las Campanas Observatory in Chile. The photometry was obtained in the I band, which is close to the standard Kron-Cousins system, and in the V band, similar to the Johnson V photometric band (Udalski et al., 2015)*. The majority of the observations were conducted in the I band filter, the specific proportion varying depending on the field. For instance, in OGLE III for the Small Magellanic Cloud, 90% of the observations were in the I band (Pawlak et al., 2013). The total number of epochs also varies depending on the field. For instance, in the I band, OGLE III exhibited a range from several dozen to approximately 3000 measurements (Soszyński et al., 2011c). In OGLE IV, the number ranged from 100 to over 750 in the I band and from several to 260 in the V band (Soszyński et al., 2015). From the information obtained for this work, we reported a cadence ranging from ~ 0.05 days to ~ 216 days, and a baseline ranging from ~ 30 days to ~ 4500 days.

From 2001 to 2009, OGLE III observations were carried on with an eight-CCD detector mosaic camera featuring a pixel scale of 0.26 arcsec/pixel and a field of view spanning 35x35 arcminutes. From 2010 to 2015, OGLE IV employed a 32-chip mosaic camera, which covered approximately 1.4 square degrees of the sky. In its final stage, OGLE achieved coverage of 3000 square degrees in the Galactic disk and bulge (Soszyński et al., 2020), along with 650 square degrees in the Magellanic Clouds (Pawlak et al., 2016), observing 70 million stars in the I band (Soszyński et al., 2023) with a magnitude range of 10-21.7 (Udalski et al., 2015).

The OGLE photometric data products are obtained according to the Image Subtraction Analysis method (Alard & Lupton, 1998) which was implemented by Wozniak (2000). The OGLE team has conducted a massive search for periodic signals in time series data. In most of their work, they used the code FNPEAKS, written by Z. Kołaczowski[†]. This algorithm employs a Discrete Fourier Transform to identify the most significant periods, incorporating amplitude and signal-to-noise ratio information. In some studies (Pawlak et al., 2013; Soszyński et al., 2016b; Iwanek et al., 2022), OGLE team combines this algorithm with others, such as the Analysis of Variance-based method

*SVO FPS Carlos Rodrigo

[†]FNPEAKS

(Schwarzenberg-Czerny, 1996), the Box-Least Squares algorithm (Kovács et al., 2002), and the Lomb-Scargle periodogram (Lomb (1976), Scargle (1982)). In Graczyk et al. (2011), instead of using FNPEAKS, the authors employed the phase dispersion minimization (Stellingwerf, 1978) and the string-length method (Lafler & Kinman, 1965).

To obtain the LCs, we started by downloading[‡] all catalogs of variable stars found in OGLE III and IV. Data were divided by mission (I, II, III and IV), field, and variability class assigned to each object in the respective field. We selected missions III and IV because of the large number of variable stars classified and the homogeneity of the classes. We downloaded three files, one for each variability class. Those files were called `ident.dat`, `variability_class.dat` (e.g., `ecl.dat` for eclipsing binary), and `phot.tar.gz`. From `ident.dat`, we obtained the identifier (ID) designated by OGLE, the right ascension, and the declination. This ID is unique for each star, although it was possible to observe the same stars in different OGLE missions, resulting in two time series for the same star. From `variability_class.dat` we obtained the ID and the periods. From `phot.tar.gz` we obtained the time series of the stars. This compressed archive contained two folders, I and V with the time series in the respective filter. We only used observations in the I band, due to its larger number of observations. We used IDs to cross-match the coordinates with the periods and time series of the stars.

We selected variability classes with the highest number of classified stars to ensure the data set was as balanced as possible and to include as many examples as possible. The selected classes include CEP, RR, M, and E, along with Long Period Variable (LPV), Ellipsoidal Binary (ELL), and Delta Scuti (DST) (for acronyms, see Column 2 of Table 2.1.1). As in OGLE IV, we also considered M and ELL as main variability classes, and extended this criterion also to the OGLE III database, in which they were originally classified as a subgroup of LPV and E, respectively.

We noted that the `ident.dat` file lists 18526 stars that did not have associated LCs in the `phot.tar.gz` file. We also attempted to download these time series directly from the online folder, but found that these stars were absent there as well. Table 2.1.1, col. 3, shows the final number of stars we successfully downloaded.

We filtered the catalog to ensure that only one time series per star was included. In instances where we had access to both OGLE III and OGLE IV LCs, we exclusively utilized those from OGLE IV. Thus guaranteeing the independence of the training, validation, and test sets (see Sect. 2.3.2). We decided not to include OGLE III LCs as independent time series because our objective was to focus on unique stars. Additionally, we did not combine time series from OGLE III and OGLE IV, as doing so could potentially introduce biases into the subset of stars with a larger observation baseline.

We aimed to ensure with high certainty that our sample contains no blended stars. In order to discard LCs coming from potentially blended sources, we adopted a very conservative threshold of three

[‡]<https://www.astrouw.edu.pl/ogle/>

elements of resolution. Meaning that we discarded sources with an additional “companion” within 1.6 arcseconds. The pixel scale of OGLE camera is 0.26 arcseconds per pixel. The median seeing, approximated by the Full Width at Half Maximum (FWHM) of the stellar Point Spread Function (PSF) measured in dense stellar fields, was about 1.25 arcseconds (Udalski et al., 2015). When using differential imaging techniques, one could obtain clean LCs for objects separated less than three times the spatial resolution. However, this assumption relied on only one of the two sources not being variable above the noise limit. Since we could not conduct dedicated inspection of close by projected companions, we adopted the previously mentioned “three elements of resolution” to guarantee that in the rest of the study we would be dealing with “non-contaminated” LCs. To identify stars that were closer than 1.6 arcseconds to each other, we employed the internal match feature of Topcat (Taylor, 2005) to search within the sample coordinates. Our goal is to identify typical variability phenomenon for each class, leading us to filter out atypical cases. The stars that were removed are listed in Table 2.1.1, column 6.

We applied sigma-clipping with a factor 3. On average, this procedure ejected approximately 0.65% of the observations from the entire time series, predominantly removing outliers. After visual inspection of a subsample of stars, the concern was raised on the impact of in homogeneity of number of observations and its impact in classification. To mitigate this possible bias, we analyzed the distribution of number of observations of our sample. “Under sampled” LCs, meaning those with less than 60 observations, only represented 1% of the first quintile were discarded, and “oversampled” LCs, those with more than 2000 observations (the 11% most sampled objects), were downsampled randomly to achieve a more uniform density distribution of observation. The typical baseline of objects with a higher density of observations was similar to that of objects with fewer observations. Subsampling objects with a greater number of observations probed the same timescales as objects in both domains. Therefore, before resampling, the minimum and maximum cadences were ~ 0.16 days and ~ 2.25 days, respectively. After resampling, the cadence was ~ 0.24 days to ~ 2.25 days.

2.3 Methodology

2.3.1 Data representation

In the phot.tar.gz files downloaded from OGLE, an original time series consisted of Heliocentric Julian Day - 2450000, the magnitude of the star, and the uncertainty of the magnitude. We used the variability period calculated by OGLE team to display the LC in phase. In this way, we visualize the variability of the star in one cycle of its period with Eq. 2.3.1.

$$\phi = \frac{t - t'}{P} - \text{int}\left[\frac{(t - t')}{P}\right] \quad (2.3.1)$$

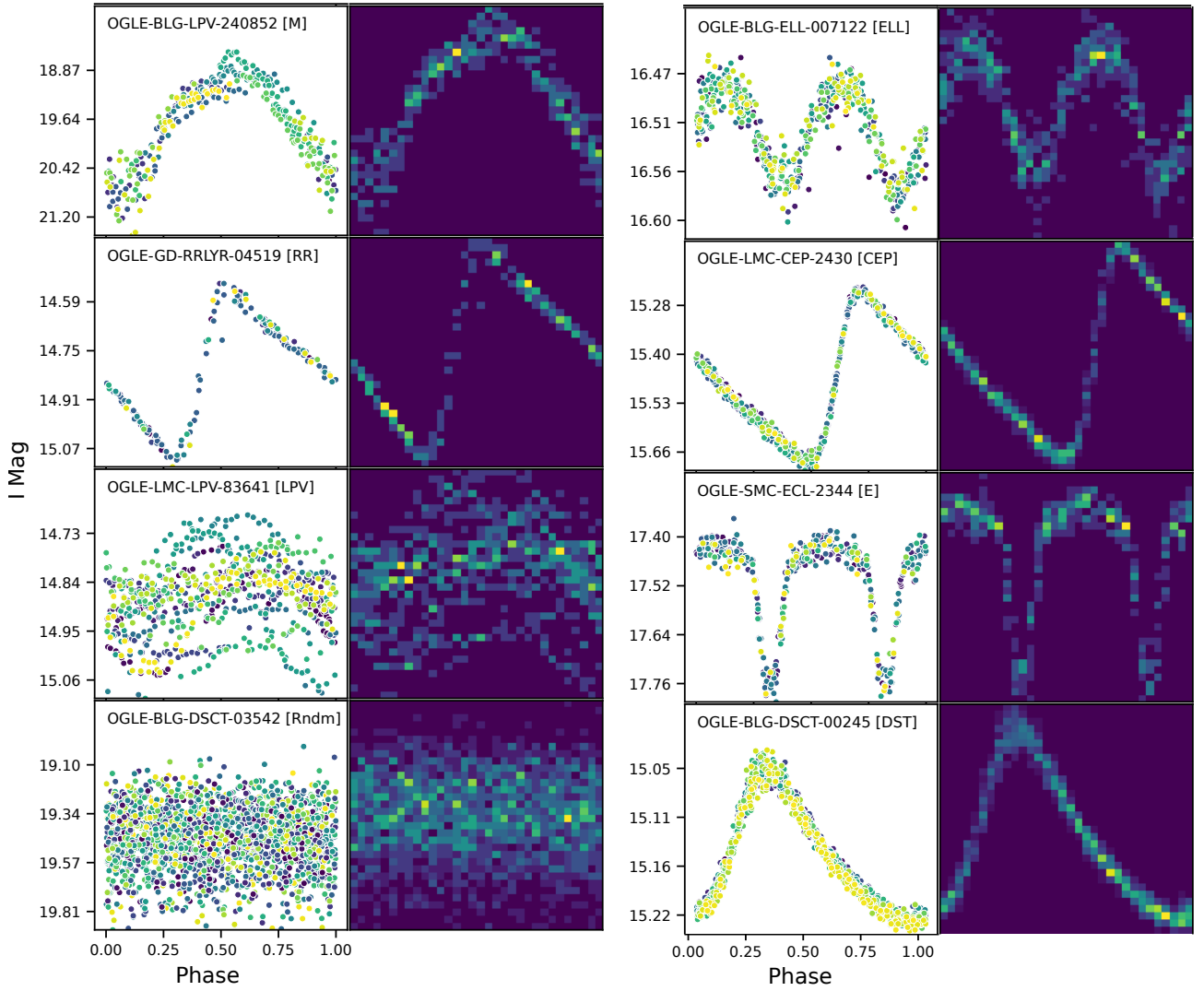


Figure 2.3.1: LCs for the different variability classes considered in the OGLE sample are presented. For each class, the first and third columns show the LCs, with colors indicating different cycles. The second and fourth columns display the corresponding 32×32 histograms, where color represents the number of observations in each bin, normalized using min-max scaling.

where ϕ is the phase, t is the time at which the measurement was made, P is the period and t' is an arbitrary epoch.

We phasefolded the LCs of the objects belonging to the seven variability classes described previously. In addition to the nominal classes in the OGLE taxonomy, we artificially populated a spurious class, by randomly resampling real LCs to arbitrary periods drawn from each class period distribution. We included this class to address a potential weakness affecting phasefolded LCs: the assumption of accurate and correct periods. If the period was inaccurate, the phasefolded image had significant scatter and showed a smooth behavior through phase. Including spurious class examples in the training set enabled the model to identify inaccurate period calculations, and, simultaneously, prevented classifying them wrongly into one of the other classes. Table 2.3.1 presented the period and amplitude domains for each variability class. We presented the first decile, D_1 , and the ninth decile, D_9 , cor-

Table 2.3.1: Period and amplitude for different variability classes.

Variability class	Field	Number of Objects	D1Period [d]	D9Period [d]	D1Amplitude [ΔI_{mag}]	D9Amplitude [ΔI_{mag}]
ELL	BLG	10723	0.46	151.1	0.07	0.28
	LMC	515	0.8	84.02	0.09	0.37
	SMC	128	0.74	10.11	0.12	0.3
M	BLG	6714	192.73	453.94	1.56	4.0
	GD	4280	226.46	490.01	1.65	4.05
	LMC	302	205.07	552.18	1.48	4.11
	SMC	70	265.73	546.74	1.76	4.34
CEP	BLG	180	0.37	14.27	0.2	0.83
	GD	1631	0.94	12.86	0.21	0.7
	LMC	4606	0.9	5.58	0.2	0.57
	SMC	4949	0.78	4.21	0.24	0.75
DST	BLG	4621	0.05	0.12	0.16	0.59
	GD	2120	0.06	0.16	0.1	0.49
	LMC	3859	0.06	0.1	0.48	1.34
	SMC	766	0.06	0.1	0.64	1.59
E	BLG	9895	0.34	4.3	0.24	1.03
	GD	300	0.28	2.77	0.16	1.28
	LMC	964	1.09	17.09	0.15	0.95
	SMC	207	0.78	26.87	0.14	0.94
LPV	BLG	7621	11.16	92.16	0.05	0.46
	LMC	3073	13.64	339.82	0.04	0.4
	SMC	672	14.26	382.68	0.05	0.35
RR	BLG	6171	0.29	0.64	0.29	0.9
	GD	918	0.3	0.65	0.31	0.94
	LMC	3669	0.32	0.65	0.43	0.96
	SMC	608	0.37	0.66	0.51	0.96

Note. Values calculated using the final preprocessed sample. We present the D1, D9, and median values for both amplitude and period. D1 corresponds to the first decile, and D9 to the ninth decile of the sample.

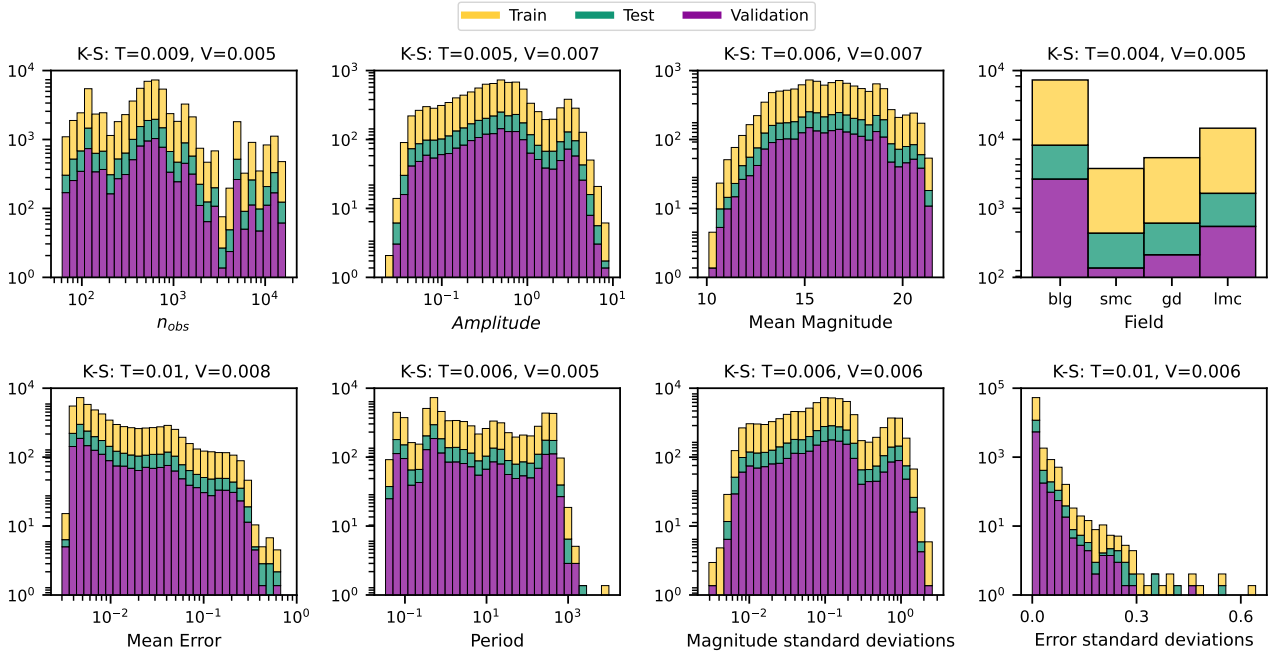


Figure 2.3.2: Distribution of observational parameters of the variable stars used in this work is shown. The number of observations, amplitude, mean magnitude, field, mean error, period, magnitude standard deviation, and the standard deviation of magnitude error are presented. The yellow, purple, and green colors represent the training, validation, and testing sets, respectively. The distributions are stacked for better visualization, and the titles correspond to the K-S test for validation and testing.

responding to the lower 10% and upper 10% of the sample, respectively. We also categorized by different environments: Galactic (Bulge and Disk) and the Magellanic Clouds (Large and Small). We were aware of variations in LCs due to environmental factors such as metallicity. Nonetheless, we aggregated different fields to create a generalized sample for our model.

We used the phase and the magnitude of the stars to represent the LC using a two-dimensional histogram of size 32x32. In this histogram, 32 bins ranging from 0 to 1 are allocated for the phase, and another 32 bins cover the range from the minimum to the maximum magnitude of each star. We then normalized the histogram to scale the bin counts to values between 0 and 1 by dividing by the histogram's maximum count. Although increasing the histogram size to 64x64 could potentially have improved classification, we chose 32x32 images to analyze more LCs with less memory, as doubling the resolution would have quadrupled the memory usage. Fig. 2.3.1 shows an example of the eight different classes of variable star selected from OGLE data.

2.3.2 Train validation and test sets

We separated the OGLE data to train the model into three sets: training, validation, and test sets in a proportion of 70:15:15. The training set was used by the algorithm to learn directly how to recognize different labeled data. The validation set was utilized indirectly to optimize the learning algorithm, aiming to quantify its level of generalization. Finally, the test set, which was entirely independent,

served to evaluate the algorithm’s classification performance, as these data had not been previously exposed to the model, either directly or indirectly.

First, we created a balanced dataset by reducing the sample size to 11 366 stars per class, matching the population of the CEP, which was the least represented. This was achieved through Undersampling (U) (see 2.3.2). We had 11 366 independent CEP, which were divided into 8 404, 1484, and 1478 to train, validate, and test, respectively. We selected the same number of examples for the rest of the classes. The total number of stars for training, validation, and testing were 67 232, 11 872, and 11 824, respectively, across eight classes: CEP, RR, L, M, ELL, DST, E, and the spurious class. For the spurious class, we selected subsets of time series from each class and chose a randomly resampled period. It is important to note that these selected time series originated from real observations, not augmented stars. Additionally, we ensured that time series from different sets were not combined. As the training set had been balanced through U, resulting in approximately 8 000 stars per class, it was hereafter referred to as ‘Train-8’.

We evaluated the representativeness of the balanced datasets by comparing the distribution of the observables and physical parameters for the samples. These comparisons are useful for characterizing the sample as astronomical objects and for assessing the representativeness of the entire algorithm’s flow. In Fig. 2.3.2, we present the distributions of various parameters across different sets. The parameters include the number of observations, amplitude, mean magnitude, field, mean error, period, magnitude standard deviations, and error magnitude standard deviations for each star. The title of each distribution show the Kolmogorov–Smirnov (K–S) test (Kolmogorov, 1933; Smirnov, 1948) for training set with validation and test. Owing to the combination of different classes and environments, the properties of each star exhibit diversity. Nevertheless, it is observable from the Fig. 2.3.2 that similar distributions prevail in the three sets for most of the distribution range. Minor discrepancies are noted only in the extremes of the magnitude standard deviations and error standard deviations.

Balanced data

In astronomy, we often encounter imbalanced datasets characterized by widely differing numbers of classified objects for several variability classes. The distribution of astrophysical phenomena in the universe is not uniform. This non-uniformity arises either from inherent biases towards intrinsically rare phenomena or from observational factors that skew detections towards specific phenomena. In the classification of variable stars, this issue becomes evident (see Table 2.1.1). E (the majority class) outnumber CEP (the minority class) by a factor of 40. This imbalance can skew standard classifiers to be overwhelmed by the larger classes and to ignore the minority class (Chawla et al., 2004). To deal with this problem, we employ different approaches to balance the data: Undersampling (U), Batch Balancing (BB), and Data Augmentation (DA).

The U technique represents a popular approach for addressing the class imbalance problem. This

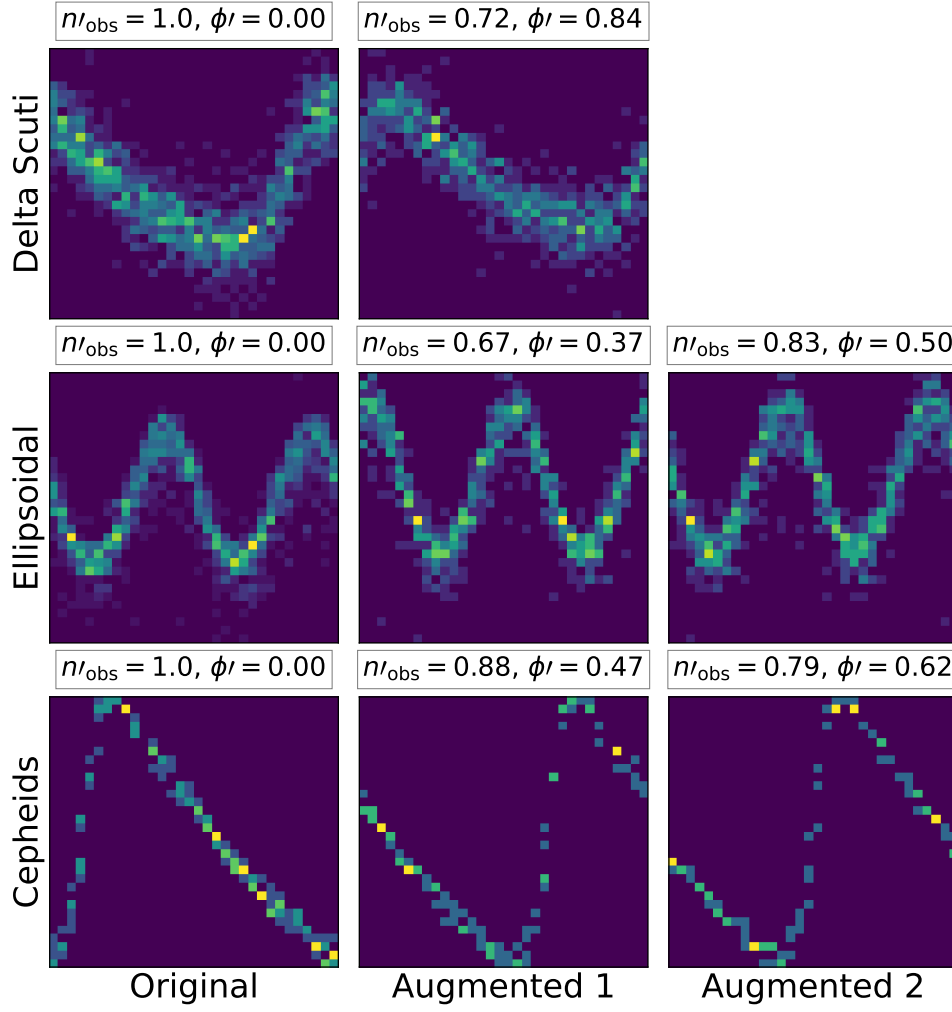


Figure 2.3.3: Examples of DA for CEP, DST, and ELL are shown. The first column represents the original LC, while the other columns represent different modifications of the LC. At the top of each image, we display the fraction of the total number of observation (n'_{obs}), and the offset in phase (ϕ').

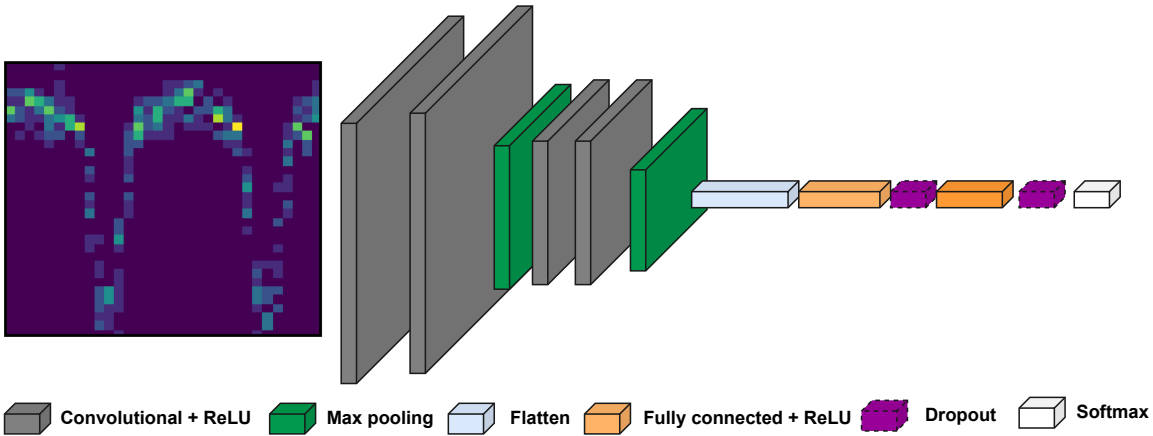


Figure 2.3.4: Representation of our convolutional neural network architecture. We show a 2D histogram of an E as an input example. Different layers of the network are displayed in various colors. The network is composed of 2 blocks of convolution layers, convolution layer, max pooling, and dropout. Then, we flatten the input and use two fully connected layers followed by the output layer. The output consists of a neuron with softmax activation. Table .2 show the hyperparameters used in this architecture.

technique involves training only with a subset of the objects populating the majority classes. The size of this subset is chosen according to the availability for training and testing of the minority classes. This approach makes it a straightforward and efficient method for handling imbalanced datasets. The main drawback is that it ignore many examples of the majority class (Liu et al., 2009). However, if we have a representative subsample of examples, we can expect favorable outcomes from this approach.

The BB involves dividing the data used to train the model into subsamples, known as batches. These batches are randomly selected without repetition from the training set 2.3.2. This procedure continues until all the data from the training set has been processed. One epoch is completed when the entire training set has been used. After completing one epoch, the entire training set is reintroduced, and the process begins again. In the BB training procedure, as described in Shimizu et al. (2018), the minority class is repeated within an epoch to ensure balanced batches.

Finally, the DA consists of generating synthetic data derived from the original data set. It is imperative that this technique is applied solely to the training set. The objective is to create synthetic data that, although different from the original, preserves relevant patterns from the data. Typical image data augmentation techniques are rotations, flips, cropping, and blurring, but they are not suitable in our context as they could inadvertently alter the intrinsic astrophysical behavior of stars (Szklenár et al., 2020). An alternative is to utilize a generative model to produce synthetic LCs, as demonstrated by (Martínez-Palomera et al., 2022) using OGLE III LCs and *Gaia* DR2 stellar parameters. However, we opted not to use this model due to potential data leakage risks. Nonetheless, their approach to DA is similar to ours. We implemented DA with three different variations in the LC: magnitude shift, phase shift, and reduction in the number of observations variations.

The magnitude shift was introduced by (Szklenár et al., 2020) who presented this methodology as a DA. This approach does not account for any correlated noise and assumes the observations are independent and identically distributed random variables, with is not necessarily the case. We complement the modifications of the LC with the other three changes. The magnitude shift consists of the generation of Gaussian noise with a mean of zero and a standard deviation equal to the observational error; then we aggregate this error to the magnitude. For phase variations, we change the value t' in equation 2.3.1. We maximize the spaced numbers of t' over the interval $0 + 1/32 < t' < 1 - 1/32$ according to the augmented LC. We created a binned LC for the reduction in the number of observations. The binning method consists of grouping data in continuous intervals and calculating a mean phase and mean magnitude per interval. We randomly chose the bin values between $0.5n_{obs}$ and $0.9n_{obs}$. For stars for which the resampling results in fewer than 60 observations, we randomly chose between 60 and the number of observations of the star.

We present an example of DA in Fig. 2.3.3. This shows DA for the three variability classes that require augmentation to achieve balanced data.

Balanced training sets

Table 2.3.2: Summary of different balanced training sets.

Training set	Star per class	Balanced technique	Minority class	Total Number
Train-8 U	8404	Undersampling	CEP	90928
Train-22 DA	22454	Data Augmentation	ELL	179632
Train-22 BB	-	Batch Balanced	ELL	165582
Train-37 DA	37045	Data Augmentation	DST	296360
Train-37 BB	-	Batch Balanced	DST	253128
Train-60 DA	60318	Data Augmentation	M	482544
Train-60 BB	-	Batch Balanced	M	369493

We created seven different training sets to investigate the limitations of the balancing techniques. Using data from Train-8 and previously unused data, we added new data to Train-8 when available. We decided the number of stars in each class based on the next minority class, aiming for using all data. For the different sets we employed BB and DA. The Table 2.3.2 presents a summary of the training set. The number in the name of the training set corresponds to the number of stars per class. The absence of information in the column stars per class is due to the lack of a fixed number of stars per class.

2.3.3 Convolutional neural network

Convolutional Neural Networks are a type of Artificial Neural Network (NN) designed to process data in the form of fixed-size multiple arrays, for example: 1d sequence, 2D images, 3d videos, among others (Lecun et al., 2015). The basic architecture of CNNs consists of four building blocks: convolution layers, regularization layers, pooling, and fully connected layers. The first three layers are used to extract the relevant features from the data, and the last layer learns the complex relationships between the features and the classes / labels. In this work, we utilized an architecture similar to that described by Szklenár et al. (2020), optimized for our input data and with fewer parameters. In this section, we provide a qualitative description of our architecture (see Fig. 2.3.4). For a comprehensive review of CNN fundamentals, we recommend Cong & Zhou (2023).

The convolution layer is the crucial component of the CNN. These layers are composed of free-learned parameters named kernels or filters of a given size. These parameters perform the convolution operation, which is a dot product between the input array and the filter, producing a feature map that summarizes the input array. During the convolution operation, we move the filter to apply the convolution across the entire array. The Padding and strides are parameters in convolution layers. Padding involves adding zero values to the array borders, while same padding adds the necessary

zero values to ensure the convolved array retains the shape of the input array. This technique prevents the loss of border information. The stride parameter determines how many steps we move the filter.

The pooling, regularization, and fully connected layers are useful for analyzing input arrays. The pooling layer aims to decrease the input size without losing essential information. Max pooling involves selecting a region in the image and keeping only the highest value. The regularization layers is used to increase the generalization of the network. The dropout is a type of regularization layers, in this technique, neurons are randomly selected and ignored during the training process (Nitish et al., 2014). The fully connected layer is a regular NN. This is a block compound with units or neurons; each unit works in parallel and is connected to all neurons in the previous layer. Each neuron received the output of all previous neurons and used this information to calculate an activated value. The activation functions are a fundamental part of the NN. It is utilized to aggregate nonlinearity to the output that enables the network to solve complex nonlinear problems. The idea is to apply some mathematical nonlinear function (δ) to the input (x) and get the nonlinear output $\delta(x)$. Many different activation function exists, and their choice translate in changes in the network's performance (Wang et al., 2020a).

In our model, the input to the CNN was the 2D histogram of the phasefolded LC. Since we used a single channel (single observational band), the dimensions of the input layer were equal to the input array of $32 \times 32 \times 1$. We incorporated two convolutional layers with 32 filters, the Rectified Linear Unit (ReLU) activation function, and same padding. We used a stride of one, meaning the filter moved one position at a time. After the two convolution layers, we applied max-pooling with a 2×2 kernel. We repeated two blocks of convolution, convolution, and Max pooling, then we aggregated two fully connected layers, regularization layers, and the output layer. We defined 2 layers with 1024 and 512 neurons. Between the fully connected layers, we inserted a dropout as regularization layers. Finally, we defined eight neurons with the softmax activation function for the output layer. The softmax function, with one neuron per class, facilitated multiclass classification. It outputted a value between 0 and 1, which represented the classification probability for each class.

This study employs a relatively standard CNN architecture, hence, its optimization is out of scope. Within this architecture, we employed the most commonly used activation functions: ReLU in the hidden layers and softmax in the output layer.

2.3.4 Training process

The training process aims at establishing a model (f) that classifies the 2D histogram of the LC (x). We define the architecture of the CNN classifier as $f_{CNN}(x, \theta)$. During training, we adjust the free parameters (θ) to create a CNN that effectively deal with the classification task: $f \approx f_{CNN}(x, \theta)$.

We employ categorical cross entropy as our loss function (J) for classification. This loss function measures the discrepancy between the predicted distribution of the model and the actual class distri-

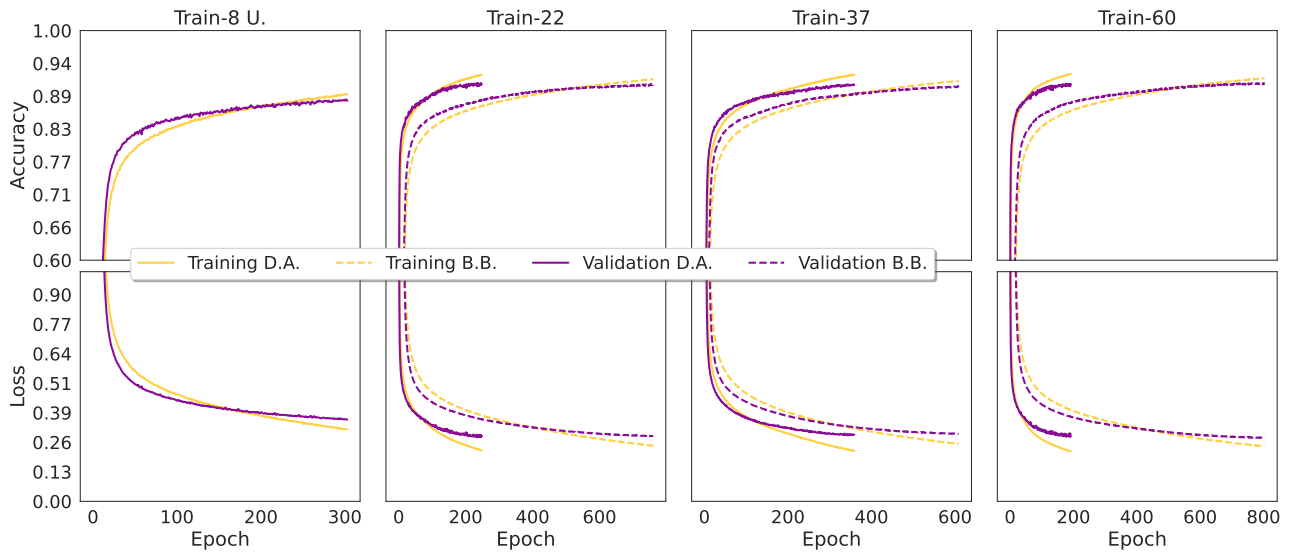


Figure 2.3.5: training process for the CNNs. The first and second rows show the accuracy and loss, respectively. The behavior of the training set is shown in yellow, while the behavior of the validation set is shown in purple. The training process is stopped at 1000 epochs or when the network starts performing better on the training set than on the validation set.

Train-8 Undersampling
F1 Score: 0.890

	ELL	M	CEP	DST	E	LPV	RR	Rndm
ELL	1375 93.0%	2 0.1%	0 0.0%	0 0.0%	96 6.5%	4 0.3%	1 0.1%	0 0.0%
M	2 0.1%	1373 92.9%	44 3.0%	9 0.6%	0 0.0%	18 1.2%	28 1.9%	4 0.3%
CEP	0 0.0%	65 4.4%	1230 83.2%	75 5.1%	0 0.0%	9 0.6%	92 6.2%	7 0.5%
DST	0 0.0%	18 1.2%	66 4.5%	1286 87.0%	0 0.0%	61 4.1%	47 3.2%	0 0.0%
E	68 4.6%	2 0.1%	0 0.0%	1 0.1%	1400 94.7%	0 0.0%	1 0.1%	6 0.4%
LPV	4 0.3%	17 1.2%	0 0.0%	79 5.3%	6 0.4%	1312 88.8%	10 0.7%	50 3.4%
RR	0 0.0%	36 2.4%	84 5.7%	95 6.4%	0 0.0%	14 0.9%	1248 84.4%	1 0.1%
Rndm	2 0.1%	69 4.7%	1 0.1%	0 0.0%	18 1.2%	85 5.8%	4 0.3%	1299 87.9%
	ELL	M	CEP	DST	E	LPV	RR	Rndm

Figure 2.3.6: Confusion matrix for the train-8 test dataset, balanced via U, illustrating initial CNN performance.

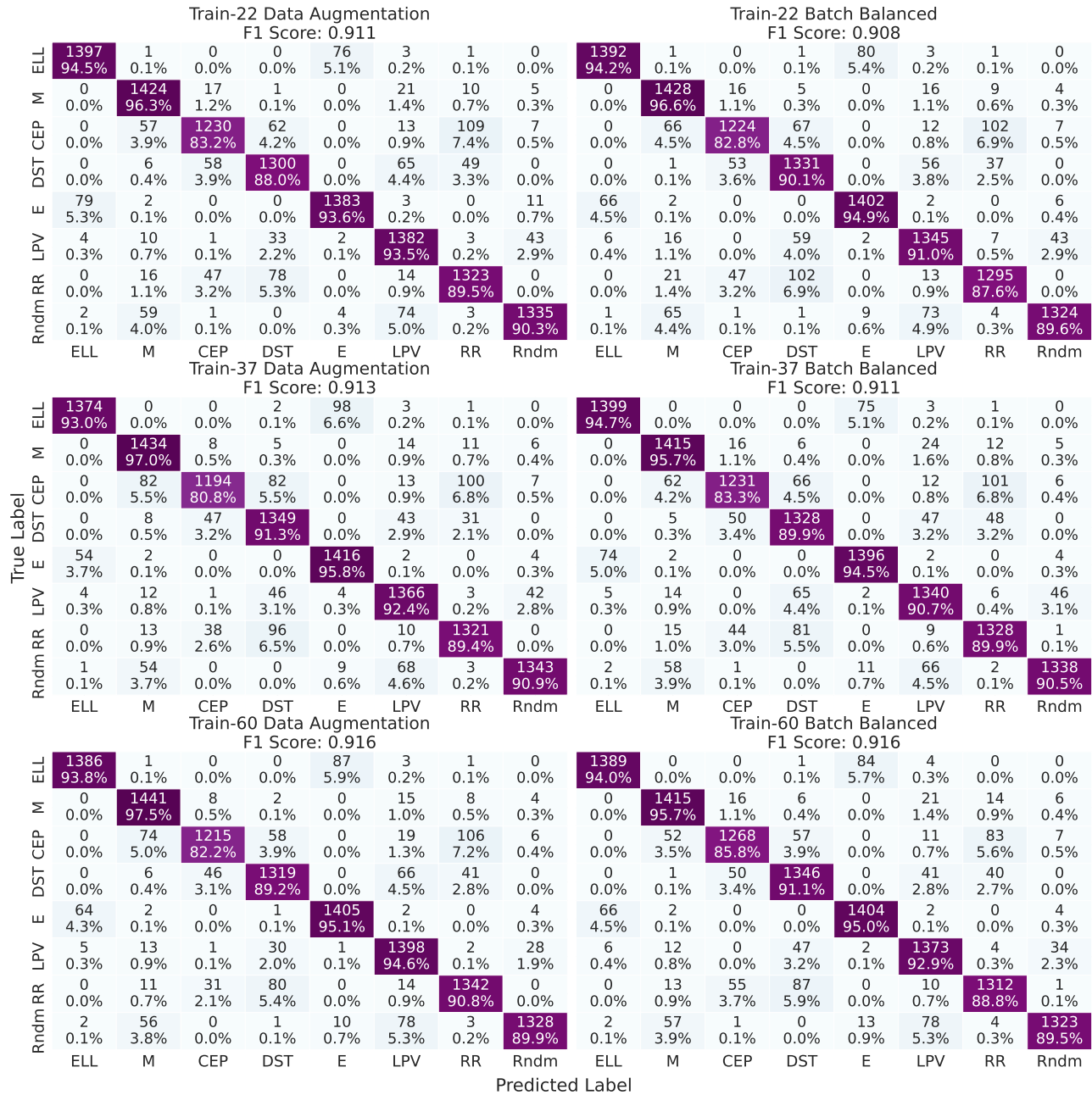


Figure 2.3.7: Confusion matrix for the CNN models in different training sets calculated in the same test set. The Y-axis is the true label obtained for OGLE, and the x-axis is the label predicted by the CNN. The boxes show the number of stars classified in each class and the percentage of the total sample.

bution. Minimizing this function narrows the gap between the prediction of the model and the true classes, thereby improving the classifier performance.

We used Adaptive Moment Estimation (Kingma & Ba, 2014) to minimize the loss function J . The parameters employed were $\eta = 0.0001$, $b_1 = 0.9$, $b_2 = 0.999$, and $\varepsilon = 0.1$. η is the learning rate and is the proportion of updating the parameters. The b_1 and b_2 are the exponential decay rate for the first and second moments, and the ε is a small constant for numerical stability (see Ruder (2016) for an in-depth discussion on gradient descent optimization).

We chose a batch size of 64 for the training sets and for the validation set. We trained the model for 1000 epochs, but stopped the training process when the algorithm showed indications of overfitting using the early stopping technique. This method actively monitors a designated quantity and ceases training when the quantity exhibits no improvement over a predefined number of epochs, termed 'patience'. For our purposes, we selected the validation loss as the quantity to monitor and established a patience threshold of 15 epochs.

2.4 Results

2.4.1 Training performance

Figure 2.3.5 illustrates the training process for the different training sets using various balancing techniques (see Table 2.3.2). The column titles indicate the name of each training set. Across the three columns, the training process for each model is depicted. The first row represents the accuracy plotted against the epoch, while the second row shows the loss against the epoch.

In Train-8 U, we present the training process for the balanced set using U. We achieved training and validation accuracies of $\sim 90\%$ and, $\sim 88\%$ respectively. The loss values for training (represented by the yellow line) began to drop slightly lower than those for validation (represented by the purple line). This pattern indicates an overfitting, where the model starts to learn the training set more effectively than the validation set. Given that these sets are independent, such overfitting can be interpreted as a halt in the improvement of the model, suggesting that further training may not be necessary.

For Train-22, Train-37 and Train-60, we display two different training processes: the solid line represents DA, while the dashed line represents BB. Both balancing approaches yielded similar results. We noted slight overfitting for the DA method and a smoother training curve for the BB technique. However, the BB technique needs more epochs to converge than the DA method.

2.4.2 Metrics

We use four well-known performance metrics to evaluate the classification model: Accuracy, Precision, Recall, and the F1 Score. For an individual class i , these scores are defined as follows:

$$\text{Accuracy}_i = \frac{\text{TP}_i + \text{TN}_i}{\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i} \quad (2.4.1)$$

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \quad (2.4.2)$$

$$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \quad (2.4.3)$$

$$\text{F1 Score}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (2.4.4)$$

Where TP_i represent the number of true positives, FP_i false positives, FN_i false negatives, and TN_i true negatives per class. Multiclass metrics are calculated as the average of multiple binary classification problems, one for each label.

To obtain a comprehensive view of a classifier's performance, it is necessary to use more than one metric. While accuracy monitors the overall performance across all classes, it does not provide detailed information about the reliability or the model's precision.

We used the seven CNN models, trained with different training sets, to obtain classification predictions for the LCs in the test set. The global performance is shown in Table 2.4.1. As we are using a balanced test set, we obtain similar values for accuracy, precision, recall, and F1 score. The Train-8 U serves as our baseline, with an F1 score of 0.891. The inclusion of more data improves the model's performance across various training sets. Train-22 DA and Train-22 BB show that DA yields better results, enhancing the model's performance. Train-37 DA and Train-37 BB result in a slight improvement, with BB providing similar performance to DA. Train-60 continues this trend, displaying similar outcomes for both BB and DA techniques.

2.4.3 Class performance

The results for each class are presented in the confusion matrices in Fig. 2.3.6 and Fig.2.3.7. The title of each figure displays the name of the training set, the balancing technique used, and the F1 score of the model. First, we present the results in our baseline Train-8 U. Subsequently, we display the outcomes for the different balancing techniques.

The most frequent misclassifications occur among pulsating stars (RR, M, DST, and CEP), which are in different evolutionary states. The amplitude and period are important parameters for distinguishing between pulsating stars. However, the algorithm does not directly access the values of period and amplitude. Instead, it indirectly accesses the periods within the pixels of the LCs, which appear much

Table 2.4.1: Performance metrics for different models.

Training Set	Accuracy	Precision	Recall	F1
Train-8 U	0.890	0.890	0.890	0.891
Train-22 DA	0.911	0.911	0.911	0.912
Train-22 BB	0.908	0.908	0.908	0.909
Train-37 DA	0.913	0.913	0.913	0.915
Train-37 BB	0.911	0.911	0.911	0.912
Train-60 DA	0.916	0.916	0.916	0.918
Train-60 BB	0.916	0.916	0.916	0.917

Note. trained on various balanced training sets and evaluated on the same balanced test set.

larger in objects with shorter periods. But, due to the different cadences, number of observations, and baselines, we obtain a certain degree of imprecision.

The other misclassifications occur among classes that denote binary star systems (E, ELL). The LC of an E results from the eclipses of the companion star. Assuming the binary does not undergo an eclipse and is close ($a \lesssim 15R_{\odot}$), the most pronounced feature in the LC emerges from ellipsoidal modulation due to tidal deformation (Green et al., 2023). This LC are typically close to sinusoidal with two equal maxima and minima in the phase (Pawlak et al., 2014). Therefore, binary systems can exhibit a combination of both eclipsing and ellipsoidal effects. This overlap of characteristics presents challenges, making accurate classification difficult.

Finally, the classification error between LPV and DST stars is not immediately apparent. DST stars are faint main sequence stars with I-band magnitudes that range from 19 to 21 magnitudes in the Large Magellanic Cloud. This puts them close to the detection limit of OGLE (Poleski et al., 2010). As a result, the noisy LCs of DST stars can resemble those of LPV.

2.4.4 Balancing strategies

Figure 2.3.6 shows the main results with the different balancing techniques. Only specific classes require balancing, CEP in Train-22, both CEP and ELL in Train-37, and CEP, ELL, DST in Train-60. CEP does not show improvement from Train-8 U to Train-60 using the DA technique, but experiences improved performance with Train-60 BB. Similarly, ELL shows only slight improvement from Train-22 to Train-37 using DA, reaching its best performance with Train-37 BB. DST follows a similar pattern, with performance decreasing from Train-37 to Train-60 DA, and the most successful results observed with Train-60 BB. These patterns indicate that the BB technique provides better performance improvement for the classes requiring balancing. Additionally, the overall good performance of DA is due to the use of more data in other classes than to an improvement in the balanced class.

Out of the seven models, Train-60 classes shows the best overall performance, and the BB give the best results for the balanced classes. We use this CNN to report the metrics in Table 2.4.2. The algorithm performs with an F1 score greater than 0.88 for all classes. ELL and E have similar values for precision, recall, and F1 score. We expect a similar distribution of False Positives and False Negatives for these misclassifications. The CEP, RR, and Random period classes have higher precision than recall. We expect more False Negatives than False Positives, meaning the algorithm is more likely to miss these classes and not identify all the stars, but the classifications made are precise. The M, DST, and LPV classes have higher recall than precision. We expect more False Positives than False Negatives, suggesting we can identify the majority of the stars in these classes, but may also find contaminants. The DST class has the lowest precision, so we expect it to be the most contaminated predicted class.

Table 2.4.2: Performance metrics obtained for the train-60 BB model

Class	Precision	Recall	F1-score
ELL	0.949	0.940	0.945
M	0.912	0.957	0.934
CEP	0.912	0.858	0.884
DST	0.872	0.911	0.891
E	0.934	0.950	0.942
LPV	0.892	0.929	0.910
RR	0.900	0.888	0.894
Rndm	0.962	0.895	0.927

2.4.5 Convolutional neural network with additional information

We compose a two-step classifier that combines the CNN with a RF that uses our convolutional model to enhance LC classification. Our CNN extracts the visual information from LCs. Additionally, we aim at integrating astrophysical knowledge, especially period and amplitude, to complement this visual insight.

We present an implementation aligned with RF due to its strong performance. RF is a supervised machine learning algorithm that functions as an ensemble of decision trees (Breiman, 2001). Each decision tree defines a potential decision path based on a subsample of the features in the training set. The final decision is the average of the decisions of each tree. We configure a RF using the default parameters of `scikit-learn` (Pedregosa et al., 2011) and with 10 trees.

We employ the Train-60 BB model to predict the variability classes for the Train-8 U, validation, and test sets, initially generating eight columns of data. These columns represent the probabilities assigned by the Train-60 BB model to each of the eight variability classes. We aggregate the period and amplitude for each star, generating tabular data comprising ten features. These features are the

input for the RF algorithm. Unlike the CNN, the RF does not require preprocessing steps such as normalization, so we do not normalize our features. We maintain consistency between the training and testing sets. In our RF analysis, we combine the validation set with the training set, as the RF method does not require a separate validation set.

Figure 2.5.1 presents the confusion matrix resulting from the RF algorithm. The combination of classifications obtained via 2D histogram phased LCs plus a few attributes enhances performance, as evidenced by improvements in the confusion matrix and the F1 score. This simple methodology demonstrates the flexibility of image classification and its synergy with other approaches. The combination of these methods enables high-precision recovery of all OGLE classes in this study.

We emphasize that we do not directly use the period and amplitude information as inputs for the CNN. The period is not explicitly included, but in a way, it is through the intensity of the pixels “in” the LC. Our goal is to develop a robust method for classifying variable stars. The main challenges include the effects of observational capabilities on amplitude variability detection and potential misclassification of periods.

2.5 Discussion

In this work we have presented an efficient methodology that combines two classifiers to codify objects in different variability classes. In the following we will discuss the astrophysical and computational performances of our methodology separately.

2.5.1 Comparison with previous works

Our research focuses on the morphological classification of variable stars using image-based classifications with phasefolded LCs. This approach is fundamentally different than other machine learning methods applied to the sequential representation of the same or similar data, for example applying Recurrent Neural Networks (RNN; Becker et al. 2020), Neural Networks (NN; Kim & Bailer-Jones 2016), Random Forests (RF; Sánchez-Sáez et al. 2021), or more recently, Transformers (Donoso-Oliva et al., 2023). In order to ease direct comparison of the results, in this section we restrict ourselves to studies based on similar taxonomy and OGLE data. However, we do compare with methodologies based not only on image representation but also “classical” sequential representation.

Image based representations

Szklenár et al. (2020) (S20), explored the use CNNs for classifying image representation phasefolded LCs from OGLE. In Szklenár et al. (2022) (S22), they broaden their methodology by incorporating a multi-input neural network into the CNN.

S20 presents and analyses a LMC dataset with 26 121 E, 24 904 RR, 2696 DST, 83 Anomalous

CNN+RF
F1 Score: 0.98

True Label	ELL	1402 94.9%	0 0.0%	2 0.1%	0 0.0%	71 4.8%	3 0.2%	0 0.0%
	M	0 0.0%	1474 99.7%	1 0.1%	0 0.0%	2 0.1%	1 0.1%	0 0.0%
	CEP	0 0.0%	0 0.0%	1465 99.1%	0 0.0%	0 0.0%	2 0.1%	11 0.7%
	DST	0 0.0%	0 0.0%	1 0.1%	1476 99.9%	0 0.0%	0 0.0%	1 0.1%
	E	84 5.7%	0 0.0%	1 0.1%	0 0.0%	1389 94.0%	3 0.2%	1 0.1%
	LPV	5 0.3%	9 0.6%	0 0.0%	0 0.0%	3 0.2%	1461 98.8%	0 0.0%
	RR	0 0.0%	0 0.0%	28 1.9%	1 0.1%	1 0.1%	0 0.0%	1448 98.0%
		ELL	M	CEP	DST	E	LPV	RR
		Predicted Label						

Figure 2.5.1: Confusion matrix for the RF classifier using the output of the CNN together with amplitude and period. The algorithm was trained using the output of the train-37. We used the same training, validation, and testing sets as in the CNN.

Table 2.5.1: F1 Score for the Variability classes Common to S20.

Class	CNN this work	S20 OGLE III	S20 OGLE IV
ACep	0.884 (CEP)	0.879	0.835
T2Cep	0.884 (CEP)	0.872	0.870
DST	0.891	0.944	—
E	0.942	0.989	0.984
RR	0.894	0.810	0.849

Note. We adopt the F1 score for CEP as the average of the F1 scores for ACep and T2Cep in S20.

CEP, and 203 Type 2 CEP. Their test set comprises 3750 augmented images for each variability type in OGLE III and 2500 augmented images for each type in OGLE IV. We compared the results of S20 with our CNN results, specifically for the variability classes that are common to both studies. Table 2.5.1 shows the comparison between the F1 score from S20 and this work. In general, their results were similar than ours; however, the creation of artificial LCs, sampled purely accounting for photometric errors (as is the strategy for data augmentation employed in S20), does not guarantee the independence of instances in the training and testing sets. Consequently, performance metrics might be artificially biased in a positive manner.

S22 expands on the work of S20 by using LCs from the Magellanic Clouds, Galactic Bulge, and Galactic Disk. In addition, the CNN results are combined with period information to classify the six main variability classes. S22 do not provide a tabular from of the metrics achieved, therefore, we compared the confusion matrices between they multiple input neural network and our combination of CNN and RF (CNN+RF). We interpret the better performance of our model for the CEP and RR classes as a result of incorporating amplitudes as features and we do not use T2 CEP and anomalous CEP. We obtained similar results for DST; this can be mostly for the period values that are smaller than the others pulsational periods and can be easier to distinguish. Finally, we have a worse performance for E, this can be explained by the fact that they are not using ELL. Overall, our results are better for the main classes, but, the inclusion of DA in the test set is a fundamental difference that makes it difficult to compare algorithms.

Time series representation

Aguirre et al. (2019) presents a deep learning method based on convolutional units to classify variable stars across multiple surveys, including OGLE, VISTA Variables in the Vía Láctea (VVV, Minniti et al. (2010)), and Convection, Rotation, and Planetary Transits (COROT, Baglin (2003); Bordé et al. (2003)). The model inputs are the differences in time and magnitude of the LC. It is trained with 8000 stars per class and survey, with a maximum of 500 observations. Additionally, a RF is trained to compare with the CNN results, extracting 59 features for each survey using the FATS library (Nun et al., 2015). Table 2.5.2 compares the accuracy reported in Aguirre et al. (2019) with our results. The combined CNN+RF model achieves better results for RR and CEP classes. We obtained marginally better accuracy for three out of the four classes in common. For E class, the lower performance is attributed to misclassification with the ELL class.

Becker et al. (2020) presents a classification model based on RNNs, which uses the differences in time and magnitude as input. This method is tested across three different surveys: OGLE-III, *Gaia* DR2 (Gaia Collaboration et al., 2018), and The Wide-field Infrared Survey Explorer (WISE, Wright et al. (2010)). The OGLE dataset incorporates a total number of 393103 stars. For comparison, a RF with 1000 trees, utilizing 59 single-band features from the FATS Library, is employed. Table 2.5.3 shows the F1 scores for the classes, compared with our work, where better results are achieved

Table 2.5.2: Accuracy of OGLE Observations as presented in Aguirre et al. (2019) compared to the results of this work.

Class	CNN+RF	Aguirre CNN	Aguirre RF
E	0.942	0.98 ± 0.01	0.97 ± 0.01
LPV	0.991	0.99 ± 0.00	0.97 ± 0.01
RR	0.990	0.94 ± 0.01	0.97 ± 0.00
CEP	0.983	0.90 ± 0.03	0.93 ± 0.01

compared to the RNN and similar outcomes to the RF. However, the processing time using FATS for the RF is ~ 7 days. Additionally, Becker et al. (2020) highlights that the OGLE LCs are biased, as they were selected from a feature-based classification, which favors these models over others. Zhang

Table 2.5.3: F1 Score of OGLE Observations as presented in Becker et al. (2020) compared to the results of this work.

Class	CNN+RF	RNN F1-score	RF F1 score
CEP	0.99	0.69	0.97
RR	0.99	0.91	0.99
DST	1.00	0.72	0.95
E	0.95	0.94	0.98
LPV	0.99	0.99	1.00

& Bloom (2021) introduce Cyclic-Permutation Invariant Neural Networks designed to be invariant to phase shifts of period-folded periodic LCs. They showcase the implementation of this neural network type using 1d Residual Neural Networks (ResNets; He et al. 2016) and Temporal Convolutional Networks (TCN; Lea et al. 2016), with the model input being the difference in phase and normalized amplitude. Utilizing data from OGLE III, they segment 163356 stars into chunks of fixed length, resulting in 540457 fixed-length LCs, and emphasize that using sequences of varying lengths can degrade accuracy. We achieve comparable results with both networks; however, we observe higher performance for E class.

The experimental design in this study differs from previous works in two main aspects: base datasets and taxonomy. While the other studies that we are comparing to used only OGLE III data (with the exception of a test set in S20), we combined all available OGLE III and OGLE IV data for objects with classes within our taxonomy. Regarding the latter, the taxonomy varies across studies. Therefore, including different subtypes of variability affect the comparisons. For example, the performance of our method for eclipsing binaries is lower than those in some of the other studies (S20 and Aguirre et al. (2019)), primarily due to the difficulty in differentiating eclipsing binaries from ellipsoidal variables (joint in a single class in the aforementioned studies).

The performance in terms of “purity” and other metrics is hard to assess since it can be affected by

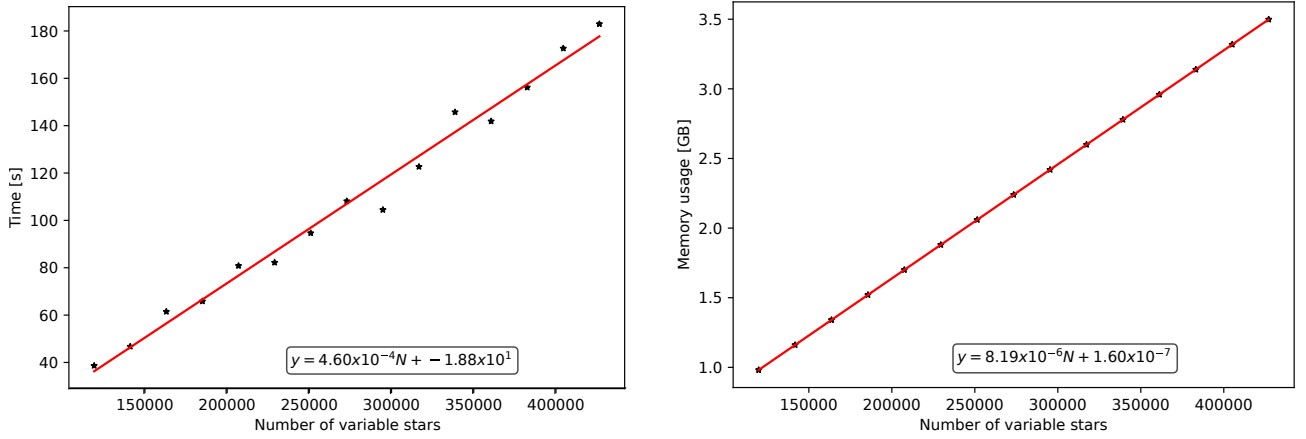


Figure 2.5.2: Computational resources required to implement the methodology of 2D histogram-based CNN for variable stars. The Figure presents the relationship between prediction time and memory usage for N variable stars represented in 2D histograms.

the different taxonomies (and, probably to a lesser extent, the base dataset themselves). However, the performance in terms of computing time is indeed comparable because we have a similar length of time series in OGLE III and OGLE IV (with a K-S test ~ 0.2), and these are directly proportional to the feature extraction time. A more direct comparison should involve using the same datasets and taxonomy, however this is not currently possible. The data links in Becker et al. (2020) and Aguirre et al. (2019) are not available, and the S20 data is not publicly accessible. In S22 and Zhang & Bloom (2021) they provided GitHub reference; But we could not find an OGLE list to identify the stars used for training and validation, making it impossible to identify independent stars for testing.

Table 2.5.4: Accuracy Score of OGLE observations as presented in Zhang & Bloom (2021) for iTCN and IResNet Models compared to the results of this work.

Class	CNN+RF	iTCN (%)	IResNet (%)
Cep	98.31	98.3 ± 0.3	98.4 ± 0.7
RRab	98.99 (RR)	99.7 ± 0.1	99.7 ± 0.4
RRc	98.99 (RR)	99.0 ± 0.2	99.1 ± 0.1
Dsct	99.73	97.6 ± 0.8	97.8 ± 0.6
EC	94.18 (E)	87.9 ± 0.9	87.8 ± 0.7
ED	94.18 (E)	95.0 ± 0.3	94.8 ± 0.4
ESD	94.18 (E)	68.7 ± 1.0	70.7 ± 0.9
Mira	99.73	97.1 ± 0.6	96.8 ± 0.3
SRV	99.12 (LPV)	96.0 ± 0.4	95.9 ± 0.2
OSARG	99.12 (LPV)	93.2 ± 0.4	93.4 ± 0.2

2.5.2 Computational resources

As previously mentioned, our goal with this project is two-fold: to produce an accurate classifier, but also an efficient one that does not involve heavy computational resources.

Our input image size is 32x32 because it allows the analysis of more LCs with less memory. If we double the size of the 2D histograms to 64x64, we could use a deeper neural network and therefore may achieve better classification results (see for example Tan & Le (2019)). However, the double size of the images quadruples the memory space they used. Therefore, it is necessary to find a trade-off between the size of the images, the complexity of the network, the performance of the model and the computational resources. We decided to give priority to the speed and efficiency maintaining a competitive performance of the algorithm.

We present two time phases: the Train Time, which includes creating histograms and train the CNN, and the Prediction Time. The Train Time is carried out on a computer with 64 GB of memory, an Intel Core i9, and an Nvidia GeForce RTX 4060 GPU. For prediction, we utilize a computer with 8 GB of memory, an Intel Core i5, and a 256 GB SSD.

Training time

Table 2.5.5 shows the time required for feature extraction and model training for a sample of 90,928 stars. The feature extraction process for the CNN is the creation of the 2D histograms, with 33% of the time dedicated to opening CSV data, 37% to phase folding and sigma clipping, and 30% to the generation of the 2D histograms. The feature extraction time for the this work RF is approximately equivalent to the CNN's time to generate the probabilities, which then serve as inputs for the RF. The RF model's training time is brief as we are utilizing only 10 trees. For comparison, we reference the time reported by Aguirre et al. (2019). which used a complete sample of 51,951 stars. It is important to note that from this sample they use a subsample as training set, alongside the use of DA, allowing for a maximum of 5 augmented stars per LC. The specifics of the training set size, however, remain undefined. In summary, our two-layered model is able to achieve results comparable

Table 2.5.5: Approximate time for feature extraction and training of algorithms.

Method	Extraction of features	Training algorithm	Total time
This work CNN	50 s	13.9 min	14.73 min
This work RF	5.44 s	1.48 s	6.92 s
Aguirre RF	11.5 days	36 min	11.52 days
Aguirre CNN	30 min	50 min	1.33 hrs

Note. We analyze a sample of 90 928 stars using 24 cores for parallel feature extraction and an RTX 4060 GPU for training. For comparison, we reference the time calculated by Aguirre et al. (2019), who utilized 6 CPUs for parallel feature extraction and a GeForce GTX 1080 Ti GPU, For a original sample of 51 951 stars.

to previous works. However, there are two main advantages to our approach: On the one hand, the identification of miscalculated periods (a “class” commonly ignored in the literature), and on the other, the computational time efficiency gain. Regarding the former, the accurate and precise identification of periods for all variability classes presents challenges, as it is dependent on both the survey characteristics and the variability classes themselves (Graham et al., 2013). Therefore, the spurious class can be used as a diagnostic on the reliability of reported periods.

The time efficiency is achieved by minimizing the number of tabular features that need to be calculated, avoiding the use of time-consuming algorithms for feature extraction, such as FATS (see Table 2.5.5). One example of a state-of-the-art approach that uses feature extraction is the Automatic Learning for the Rapid Classification of Events (ALeRCE, Förster et al. (2021)). It processes alerts from ZTF, which contain photometry in g and r bands. It employs a Balanced Random Forest (Chen et al., 2004) with 500 trees and 152 recent and optimized feature extraction packages (Sánchez-Sáez et al., 2021). A fair comparison with their tools in production is not possible because the data types are different (they process ZTF Avro alert files 4[§]). Furthermore, the LC classifier of ALeRCE tackles a more general problem than the one addressed here. But, since their architecture and procedures are public, we conducted a comparison experiment: We selected randomly balanced samples of stars with different amount of epochs of observations (70 stars per bin on the numbers of epochs as shown in Table 2.5.6). For the performance comparison, we focused on 61 single-band features from the ALeRCE pipeline (61-AF) and timed the feature extraction plus classification flow using a 500-tree forest. These figures (computing time) were contrasted with those from our approach: feature extraction and classification with 10 trees. Table 2.5.6 shows the results of the comparison. Both experiments were conducted without parallelization and in the same personal laptop. As can be seen (and somewhat expected intuitively), the feature extraction time increases significantly with the number of observations, whereas our method maintains a constant time regardless of the number of observations. This experiment confirms the intuitive idea that our methodology computational requirements scales (in the sense of actually not scaling) nicely with density / size of the lightcurves and hence it is a good alternative to traditional methods. However, we must emphasize in the methodologies as future facilities will provide both, dense and sparse data sets and will benefit from focused / smaller and wider taxonomies.

In a forthcoming paper, we aim to investigate the synergy between our method and various periodograms to identify the optimal combination of methods for accurately recovering the variability class with a reliable period.

Predicting time

We tested the algorithm on a 8GB of memory, Intel Core i5, and a 256 GB SSD. In Fig. 2.5.2 we show the time and memory used to classify different samples of stars.

[§]<https://zwickytransientfacility.github.io/ztf-avro-alert/>

Table 2.5.6: Classification time for 70 stars per number of observation bin.

n_{obs}	61-AF+RF [s]	CNN+RF [s]
(60, 100]	0.142 ± 0.042	0.034 ± 0.001
(100, 500]	0.486 ± 0.284	0.034 ± 0.002
(500, 800]	1.337 ± 0.385	0.034 ± 0.001
(800, 1500]	2.696 ± 0.579	0.034 ± 0.001
(1500, 2000]	4.533 ± 1.262	0.034 ± 0.001

Note. We show the time for 61 single-band features from the ALeRCE pipeline and compare it with the classification time of CNN+RF.

We can classify and represent half a million stars in 3 minutes and using 3.5 GB of memory, respectively. We do not consider the time required to create a 2D histogram. However, we have made the complete OGLE catalog used in this work publicly available, along with the corresponding 2D histogram for the OGLE data, on GitHub[¶].

These results are promising and competitive in terms of speed and accuracy. As (Catelan, 2023) indicates, the *Vera Rubin* Observatory is expected to detect up to 10^8 variable stars. Our algorithm is potentially able to process this volume of data in approximately 13 hours, but it may not be feasible to handle such a quantity on a standard computer. Therefore, further research is necessary to optimize data representation and analysis for more accessible processing. This big data scenario marks a significant paradigm shift in scientific research methodology. Consequently, there is a pressing need to develop tools that are accessible to the broader community of astronomers, particularly those with limited access to high computational resources, with the ultimate goal of democratizing science.

2.6 Summary and conclusions

In this study, we have introduced a methodology to classify variable stars based on the morphology of their LCs. We present a 2D histogram, sized 32x32, to depict LCs as images. We introduce a Convolutional Neural, consisting of convolutional layers, max pooling, and dense layers.

We select our variable stars data from the OGLE. We choose variability classes with a reasonable number of examples per class. Using these stars, we select eight classes: RR, CEP, LPV, M, ELL, DST, E, and an additional artificial class representing LCs with misclassified periods.

We applied three distinct approaches to manage the unbalanced dataset: DA, BB, and U. BB emerged as the most effective, achieving an F1 score of ~ 0.92 . In contrast, DA led to less favorable outcomes in the augmented classes compared to those obtained with U. Our DA strategy includes phase shifts, random reductions in the number of LC observations, and magnitude shifts within error margins. The

[¶]<https://github.com/Monsalves-Gonzalez-N/Paper.OGLE>

suboptimal performance of DA suggests that these techniques may not reliably provide novel and beneficial information to the algorithm.

We achieve a limit in the image classification of $\sim 92\%$. We justify this value because it is challenging to differentiate pulsating stars without period and amplitude values. Thus, we test a two step algorithm with a RF that incorporates the neural network output along with period and amplitude values. With this adjustment, the convolutional network identifies almost all variability classes, except for distinguishing between Ellipsoidal and Eclipsing variations, where a 5% discrepancy remains.

Finally, we outline the resources necessary for implementing our proposed methodology, emphasizing the efficiency of our approach. Our method demonstrates notable efficiency: opening histograms for variable stars requires approximately 4 gigabytes of memory for half a million variable stars. Moreover, the classification process takes around 180 seconds for this volume of data on a standard computer. On a better computer, we are capable of creating approximately 76 images per second per core and a prediction time of $\sim 60\mu s$ per star. Such efficiency is noteworthy and contributes to addressing the challenges of big data in astronomy.

Chapter 3

Astrophotometric search for massive stars in the Milky way. Confronting Random Forest predictions with available spectroscopy.

This chapter is recommended for publication pending minor revisions, as:

Astrophotometric search for massive stars in the Milky way. Confronting Random Forest predictions with available spectroscopy.

N. Monsalves, A. Bayo, M. Jaque Arancibia, J. Bodensteiner, A. G. Caneppea, P. Sánchez-Sáez, R. Angeloni. *Astronomy & Astrophysics*

Abstract

Massive stars play a significant role in different branches of astronomy, from shaping the processes of star and planet formation to influencing the evolution and chemical enrichment of the distant universe. Despite their high astrophysical significance, these objects are rare and difficult to detect. With Gaia's advent, we now possess extensive kinematic and photometric data for a significant portion of the Galaxy that can unveil, among others, new populations of massive star candidates. In order to produce bonafide bright (G magnitude ≤ 12) massive star candidate lists (threshold set to spectral type B2 or earlier but slight changes in this threshold also explored) in the Milky Way subject to be followed up by future massive spectroscopic surveys, we have developed a Gaia DR3 plus literature data based methodology. We trained a Balanced Random Forest (BRF) with the spectral types from the compilation by Skiff (2014) as labels. Our approach yields a completeness of $\sim 80\%$ and a purity ranging from 0.51 ± 0.02 for probabilities between 0.6 and 0.7, up to 0.85 ± 0.05 for the 0.9–1.0 range. To externally validate our methodology, we searched for and analyzed archival spectra of moderate to high probability ($p \geq 0.6$) candidates that are not contained in our catalog of labels. Our independent

spectral validation confirms the expected performance of the BRF, spectroscopically classifying 300 stars as B3 or earlier, including 107 new stars. Based on the most conservative yields of our methodology, our candidate list could increase the number of bright massive stars by $\sim 50\%$. As a byproduct, we developed an automatic methodology for spectral typing optimized for LAMOST spectra, based on line detection and characterization guiding a decision path.

3.1 Introduction

It is commonly agreed that the galactic stellar mass distribution peaks at $\sim 0.5M_{\odot}$, and that its shape has a certain universality (Bastian et al., 2010). However, the slopes towards the higher and lower mass ends are still under debate. Despite their lower numbers in our Galaxy, the effect of massive stars ($M > 8M_{\odot}$) on their surroundings is critical. These stars play a fundamental role in astrophysics, from shaping the processes of star and planet formation (Bally & Zinnecker, 2005) to providing us with key knowledge of the distant universe (Eldridge & Stanway, 2022). In terms of direct outcomes, massive stars can produce a wide range of phenomena, such as Type II supernovae (Smartt, 2009), neutron stars and black Holes (Heger et al., 2003), gamma-ray bursts (Izzard et al., 2004; Nakar, 2007), and due to their high multiplicity fraction (Sana et al., 2012; Moe & Di Stefano, 2017; Barbá et al., 2017), they are sources of gravitational waves (Schneider et al., 2001).

Observational classification of stars regarding their mass cannot be directly achieved (as the mass is not an observable). However, classification systems like the MK (Morgan & Keenan, 1973) aim to map observables (spectral types) to masses. This system is anchored with a set of standard stars, and classification relies on comparing spectral morphology with these standards. The system focuses on the blue part of the optical spectrum, which historically were the regions of highest sensitivity in photograph plates (Pickering, 1890; Walborn, 1972; Gray & Corbally, 2009; Sota et al., 2011, 2014). Regarding massive stars, the $8 M_{\odot}$ boundary corresponds to spectral type B2 for dwarfs (V), B5 for giants (III), and B9 for supergiants (Ia/Ib, Morgan 1951; Gray & Corbally 2009; Maíz Apellániz et al. 2024).

Focusing mostly on the Milky Way, extensive efforts have been made to identify and spectroscopically classify massive stars. These include, among others, targeted studies (Sota et al., 2011; Simón-Díaz et al., 2011b,a; Maíz Apellániz et al., 2013; Sota et al., 2014; Simón-Díaz et al., 2015; Li, 2021), specialized compilations such as Pantaleoni González et al. (2021), and broader spectral type compilations (Skiff, 2014). Combined, these works yield a total of ~ 7500 stars spectroscopically classified as B2 or earlier.

This low figure has been subject of a heated debate for several decades. Kroupa & Weidner (2003) reported an apparent lack of massive stars in the field when compared to their relative number with respect to late-type stars in clusters. This finding is supported by observational studies such as Rybizki & Just (2015), however, due to the low number and non-homogeneous distribution of high-mass

stars, the slope of the mass function above $10 M_{\odot}$ is an extrapolation. One possible explanation for the apparent deficit of field massive stars is provided by the Integrated Galaxy initial mass function theory (Kroupa et al., 2024), but see Cerviño et al. (2013a), and Cerviño et al. (2013b) for statistical counterarguments. It is therefore necessary to obtain more complete and robust censuses of high-mass stars.

Astronomy is undergoing a transformative phase with an unprecedented influx of data from large-scale surveys. The Gaia mission with its Data Release 3 (Gaia DR3, Gaia Collaboration et al. 2023) has delivered six-parameter astrometric data for nearly 900 million stars. This information will soon be complemented by upcoming spectroscopic surveys, for instance, the 4-metre Multi-Object Spectroscopic Telescope (4MOST) (de Jong et al., 2016; de Jong et al., 2019), WHT Enhanced Area Velocity Explorer (WEAVE) (Jin et al., 2024), Sloan Digital Sky Survey-V (SDSS-V) (Kollmeier et al., 2019). Additionally, VLT/MOONS (Gonzalez et al., 2022) will be equipped with 1000 fibers, significantly more than any existing near-infrared spectrograph. Therefore, to fully and efficiently exploit these extensive capabilities, it is essential to inform the surveys with the most robust candidate lists.

Within this context, Zari et al. (2021) presented a comprehensive catalog of hot, luminous stellar candidates with estimated spectral types from O to A in the Milky Way. An immediate goal of the work was to provide robust candidates for SDSS-V follow-up (Kounkel et al., 2023). The candidate catalog consisted of ~ 990000 sources with $G < 16$ mag selected with Gaia Early Data Release 3 (Gaia eDR3; Gaia Collaboration et al. 2021) photometry and astrometry and Two Micron All Sky Survey (2MASS; Skrutskie et al. 2006) photometry. The sample was designed to be more complete than pure with a reported purity of 20% for stars with temperatures consistent with spectral type B7V.

In this work, we present a supervised machine learning approach designed to improve the yield of massive star candidates with spectral types B2 or earlier in the Milky Way. The method builds on and updated version (fully Gaia DR3) the selection criteria proposed by Zari et al. (2021). In order to validate our approach, we mined spectroscopic archival data for our candidates. As a byproduct, we derived a decision path for the classification of low-resolution spectra of massive star candidates. The path relies on the equivalent width and the presence or absence of specific spectral lines.

This paper is organized as follows: in Section 4.2, we describe the parental catalog along with its characteristics, and we describe the sources of spectroscopic labels. In Section 3.3, we explain the methodology employed to identify moderate to high probability candidates. In Section 3.4 we present the BRF results and the moderate to high probability candidates of spectral type B2 or earlier. In Section 3.5, we discuss the validation of our methodology, including external spectroscopic diagnostics and the comparison with databases and catalogs specialized to massive stars. Finally, in Section 4.5, we present the conclusions and summarize the main results of this work.

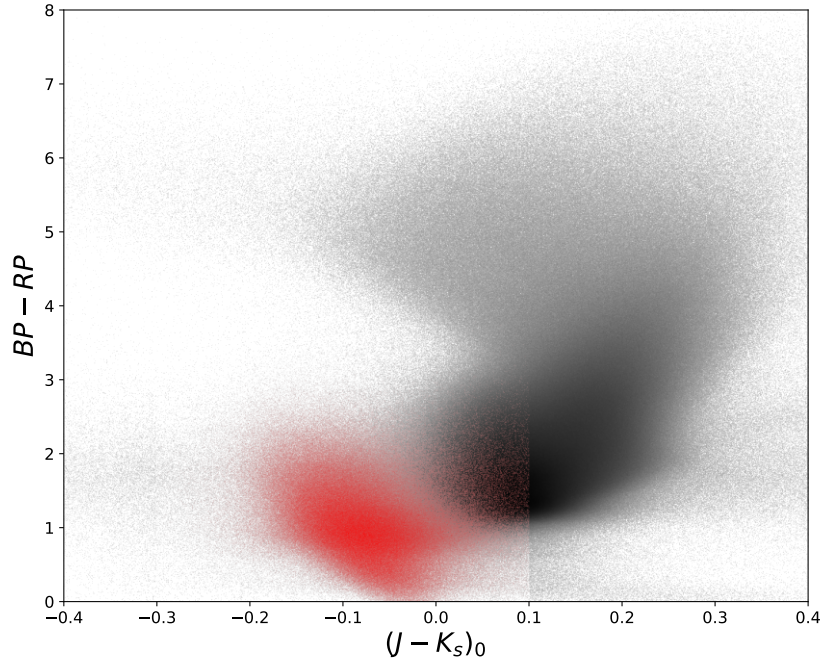


Figure 3.1.1: Recreation of the color-color diagram of Zari et al. (2021) adapted to the full DR3 from Gaia. The entire population brighter than $G_{\text{mag}} < 16$ is shown in black, while the OBA GDR3 astrophotometric sample is depicted in red. In this plot, it is easy to distinguish that the brighter stars occupy specific positions in the color-color diagram.

3.2 The data

In the following we present the different sets of data gathered to develop our classification flow, namely, the source of attributes and labels.

3.2.1 OBA GDR3: Parental catalog

The Zari et al. (2021) catalog consisted of sources in Gaia Data Release 2 (Gaia DR2, Gaia Collaboration et al. 2018) crossmatched with 2MASS (see Section 2.1 of Zari et al. 2021). Posterior to the matching, the authors updated the Gaia photometry and astrometry to those of the Gaia Early Data Release 3 (Gaia eDR3, Gaia Collaboration et al. 2021). Their selection criteria filtered stars brighter than $G < 16$ mag, and then applied simple photometric cuts to select stars of approximate spectral type OBA.

To define our parental catalog, we took advantage of the availability of Gaia DR3 and the existing Gaia DR3–2MASS crossmatch*. We applied their selection function to this updated release. From now on, we will refer to this catalog as OBA GDR3, which contains 1016743 entries, $\sim 3\%$ more than OBA GR2. This difference maps naturally to the difference in “seed” catalog as Gaia DR3, as a whole, contains $\sim 7\%$ more sources than Gaia DR2.

*ESAC’s external crossmatch catalog service

Figure 3.1.1 presents a color-color diagram of the entire $G < 16$ mag population in Gaia DR3, together with the OBA GDR3 subsample corresponding to a slightly updated version of Fig. 2 in Zari et al. (2021).

3.2.2 OBA GDR3: Features description

We used photometric, astrometric, and color information as features in our analysis. Regarding photometry, the Gaia bands (G , G_{BP} , and G_{RP}) covered the optical domain and the 2MASS ones (J , H , and K_s), the near-infrared. In addition, we considered 15 color indexes resulting from the combination (differences) of pairs of photometric bands. The astrometric parameters considered were parallax (ϖ), proper motion in right ascension (μ_α) and declination (μ_δ) with their respective errors (σ_ϖ , σ_{μ_α} , σ_{μ_δ}).

Finally, we incorporated six quality indicators that measure the goodness of fit of the astrometric solution, such as the `astrometric_sigma5d_max` (ASM), `astrometric_excess_noise` (AEN), Renormalised Unit Weight Error (RUWE), `astrometric_gof_al` (AGOFAL), `ipd_gof_harmonic_amplitude` (IPDGOFHA), and `astrometric_excess_noise_sig` (AENS). Extreme values of these features have been interpreted in the literature as signs of multiplicity (Belokurov et al., 2020; Dodd et al., 2024). For more detailed information on these parameters, refer to the Gaia Data Release 3 documentation.[†]

3.2.3 OBA GDR3: Spectral type compilation

The aim of this work is to identify reliable early type candidates based on the features presented in the previous section. We tackle this problem as a supervised classification one with a taxonomy based on spectral types. Obviously, not all objects in the OBA DR3 catalog have spectral type determinations, but a subset do and those will compose the “universal truth” that we want to propagate with our methodology. The spectral types will be referred to as *labels* from now on, and we used two sources for them: the spectroscopic catalog from Skiff (2014), hereafter referred to as Skiff and wider and very commonly used database, Simbad (Wenger et al., 2000). We note that this is a manually curated survey that may include mismatches and fall behind the current literature; however, it is the most complete stellar database available at present.

To assign labels to the entries of the OBA GDR3 catalog, we performed two independent cross-matches: to Simbad and Skiff. In both cases, we keep every entry from OBA GDR3 and employed a 2radius for coordinate matching.

Regarding multiple counterparts, over 95% of targets in Simbad had a one to one correspondence in the defined radius, and hence, in case of multiple-match, we opted to use as “true” counterpart, the

[†]GDR3 Documentation

closest target in terms of angular distance. The situation was drastically different in the matching with Skiff, where over 70% of the targets have indeed multiple counterparts. This was expected, as Skiff, by construction, does not prevent multiple entries for the same astrophysical object, in Sec. 3.2.3 we present our approach to deal with duplications.

Numerical encoding

The labels provided in Skiff, and Simbad are alphanumeric quantities that we proceeded to encode in numerical tuples to preserve the ordered relation of the labels (an O stars is closer in properties to a B star than to an A and so forth). These encodings are used in Section 3.3 to train the model and revised in Section 3.5.3 to assess possible limitations.

For the first element of the tuple, the letter, O to M, was mapped to integer values in increments of 10, ranging from 0 to 60. If a subtype was reported in either catalog, its integer value (0–9) was added. Otherwise, in cases where a subtype was not present, an integer was sampled from a uniform distribution of possible values and then added. For instance, B4 was encoded as 14.

In addition, whenever uncertain values or ranges (e.g., B9/A0, F1-F3) were reported, the individual encoding for the extremes were computed and used as the inclusive limits for a distribution to be uniformly sampled. It must be noted that this sampling artificially “produces” spectral types for which no spectra of standard stars may exist (Maíz Apellániz et al., 2024), however, the classifier (presented in Sec. 3.3) will be trained on a lightened binary taxonomy “earlier or later than B2”, and hence this (ease of computation motivated) artifact can be neglected. Finally, when the uncertainty in the spectral classification spanned more than one main spectral type (e.g., OB), the label was saved as a known label but excluded from the training, while targets with reported composite spectral types (such as kA2hA9mF2[‡]) were discarded.

The second element of the encoded tuple captured luminosity class information for the targets not discarded in the previous step. Each luminosity class, from hypergiants to subdwarfs, was assigned an integer value from 0 to 5 (as hypergiants and supergiants were grouped with the value of 0). For intermediate classes (e.g. II-III, II/III) the mean value, rounded to one decimal was adopted (for instance 2.5 was adopted for III/IV). In addition to the tuple, reported spectral peculiarity information was propagated by means of five, not mutually exclusive, flags (*uncertain*, *miscellaneous*, *line-broadening*, *emission*, and *nitrogen/carbon enriched*, see Table .3)).

The spectral type encoded labels were further simplified into earlier than a given threshold or not (see Section 3.2.3); and the encoded luminosity classes were kept aside to serve as a posteriori diagnostics to analyze trends in the predictions.

[‡]This corresponds to A2 for the Ca K-line subtype (k), A9 for the hydrogen-line type (h), and F2 for the metallic-line type (m)

Since the different elements used as attributes among themselves (photometry and kinematics) and labels are not coetaneous, for objects exhibiting extreme variability this could pose some source of noise in the modeling. However, we expect the number of such cases to be negligible. For example, the fraction of extreme variability cases reported in the VISTA Variables in the Via Lactea (VVV) survey (Minniti et al., 2010) is below $10^{-5}\%$ (?).

Addressing repeated entries

While the (2radius) crossmatch of OBA GDR3 data with Skiff may yield multiple entries for the same astronomical sources, either due to different sources or to multiple spectral classifications of the same source, the crossmatch with SIMBAD does not. Thus, in order to faithfully compare both crossmatched datasets against each other in the following subsection, we first performed an internal match within the Skiff-correlated set with respect to its astrometric coordinates, as to identify and group multiple records that describe the same object. The match threshold criterion was a maximum radius of 2of angular distance. For each group, we computed the means and standard deviations of their spectral type and luminosity class encodings. Specifically, grouped entries with standard deviation values greater than 1.0, 3.0, and 5.0 (i.e., 1, 3, and 5 units of spectral subtype) were discarded, resulting in three filtered sets. This standard deviation will hereafter be referred to as $\sigma_{\text{literature}}$.

Comparing the two sources of labels

We compared Simbad and Skiff in detail to assess the differences between these two databases. Fig. 3.2.1 shows the number of spectral types recorded in Skiff and Simbad, considering the previously discussed $\sigma_{\text{literature}}$. We collected spectral type information for approximately 55000 stars, which represented 5% of the OBA GDR3 sample. In all cases, Simbad contains more stars with spectral type information than Skiff. For O stars in Skiff, the number increases by 30% when the $\sigma_{\text{literature}}$ threshold rises from 1 to 3 and by 5% when it increases from 3 to 5, with a similar trend observed for other spectral types. Based on this, we adopt a $3 - \sigma_{\text{literature}}$ threshold for grouping Skiff data, achieving a balance between robustness and sample size.

Figure 3.2.2 compares the luminosity classes reported in Skiff and Simbad, with Skiff containing slightly more stars with reported classifications. Both databases exhibit a similar distribution, with dwarfs being the most abundant class. However, there is a discrepancy in the number of subgiants (class IV), due to uncertainties in the luminosity classifications in Skiff (e.g., IV/V or V/IV), which are encoded as intermediate between these classes (see Section 3.2.3). When combined with the binning process used in the histogram, these uncertainties result in a higher contribution for subgiants.

Figure 3.2.3 compares the labels identified for the same object in the two databases for those cases where the luminosity class is in agreement. We included various dispersion metrics for comparison: the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Coefficient of Determination (R^2), Pearson Correlation Coefficient (r), and Spearman's Rank Correlation Coefficient (ρ), available

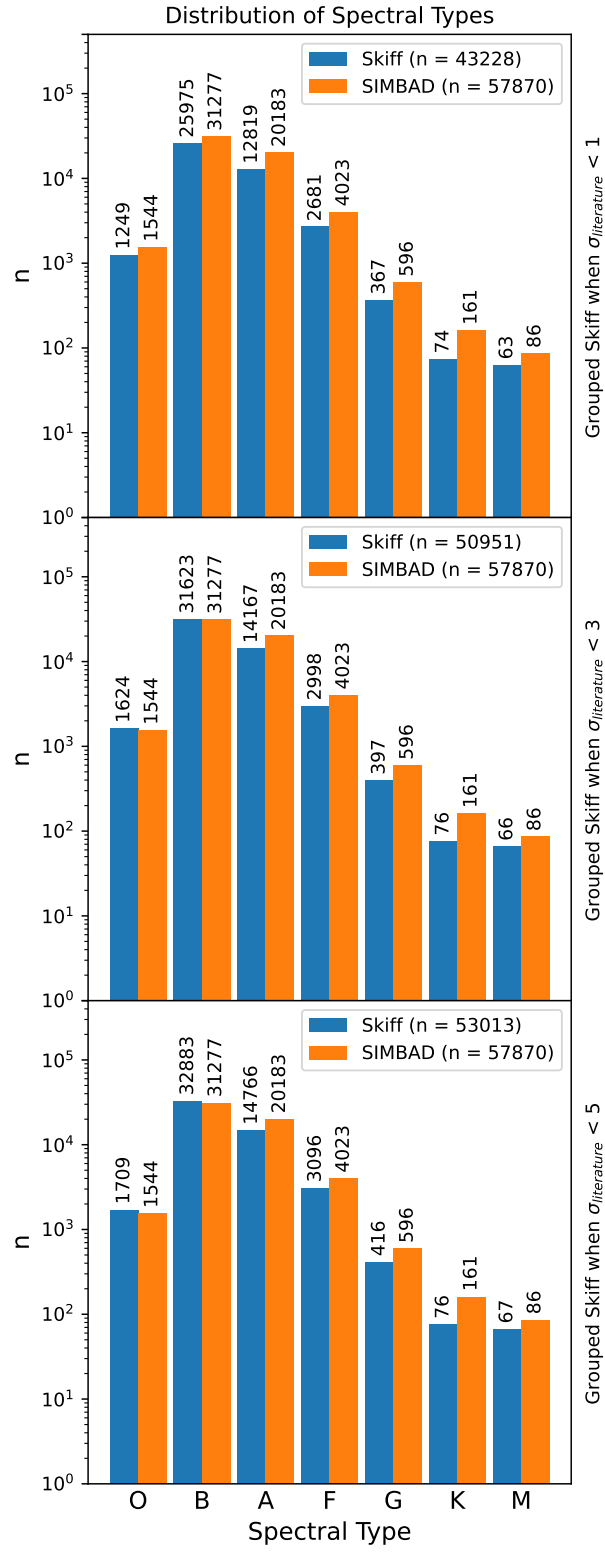


Figure 3.2.1: Distribution of OBAFGKM spectral types for Skiff and Simbad. The subplots from top to bottom correspond to different $\sigma_{\text{literature}}$ thresholds for Skiff (see text).

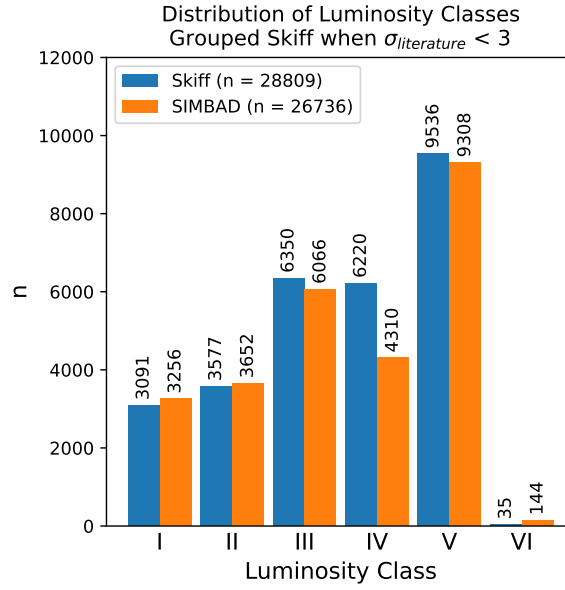


Figure 3.2.2: Distribution of luminosity classes for Skiff and Simbad. The luminosity class reported for Skiff is based on a $\sigma_{\text{literature}}$ threshold of 3.

in Pedregosa et al. (2011). The MAE and RMSE were 0.36 and 0.91, respectively, with differences remaining below one spectral subtype unit. The other metrics were very close to 1, confirming a strong correlation between the two databases and indicating that they were highly comparable. We identified a few outliers above the bulk of the data and investigated them using both the SIMBAD and Skiff databases. In most cases, we could not retrieve the reference listed in SIMBAD. In a few others, we found misreferences in Skiff or more updated spectral classifications in one of the two databases.

Given this strong correlation, we ultimately decided to use the Skiff database for our analysis. This choice is motivated because approximately ~ 1000 objects in Simbad lack references for their spectral types; ~ 7000 objects potentially suffer from suboptimal astrometry, with a positional precision of 8–12 compared to the ~ 1 coordinate precision guarantee in Skiff; and, finally, Skiff ensures spectroscopic origins for its compilation.

Simplification of encoded labels and adopted threshold

We simplified the encoded labels into more global and representative categories. Since the fundamental truth sample is meant to provide 'standard examples' of each class, we filtered out phase-dependent spectral types (e.g., OBe), stellar spectra with reported uncertainties in the classification, and non-MK classifications (e.g., Wolf-Rayet stars, white dwarfs, planetary nebulae, and carbon stars). Additionally, 'em' designations were removed, as they lack information about the primary spectral type. Lastly, we excluded the ~ 2000 systems reported in the catalog as multiples, due to the limited number of cases (only 2%) in which spectral types for both components are provided. This decision could be seen as at odds with the fact that multiplicity is known to scale with primary mass. However, the uncertainty introduced by the lack of individual spectral types (and hence across the period - mass-ratio

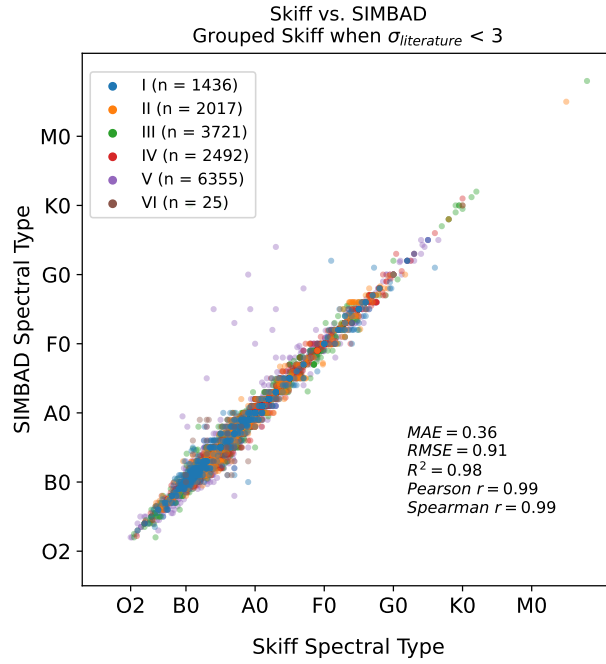


Figure 3.2.3: Comparison of spectral types for Skiff and Simbad. The different colors represent luminosity classes. Only luminosity classes that coincide in both databases were considered.

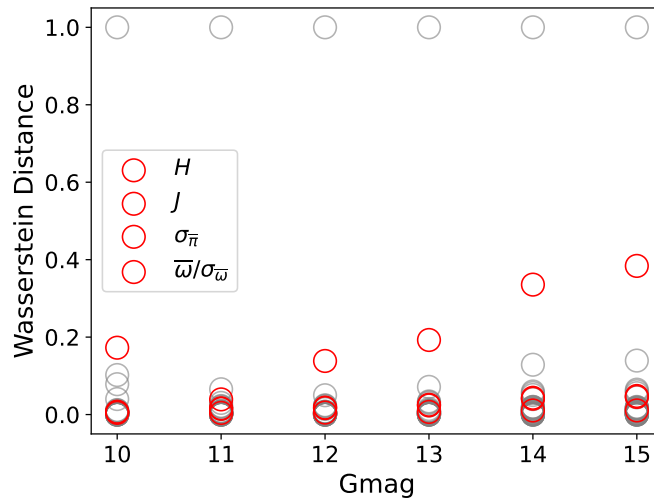


Figure 3.2.4: Normalized Wasserstein distance between the OBA encoded labels and the pool of new candidates for each feature for different G band magnitude. Each feature is shown in grey, while the four most important features for the model (see Section 3.4) are highlighted in red.

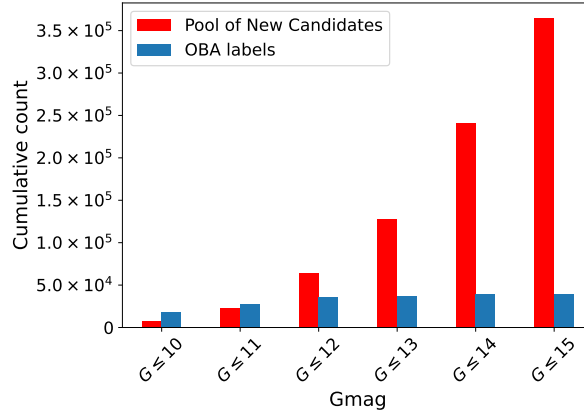


Figure 3.2.5: Cumulative number of sources for different G band magnitude cuts for the OBA encoded labels and the pool of new candidates.

parameter space) for these systems was deemed detrimental to the training process.

The ”boundary” between massive and not massive stars is not a simple cut in terms of spectral type for B stars, as the luminosity class has to be considered. However, more than half of the available encoded labels lack this information. Promoting purity over completeness, we set B2 as the threshold, as this spectral type corresponds to a massive star across all luminosity classes (Maíz Apellániz et al., 2024).

Table 3.2.1: Number of classifications excluded

Spectral type (labels)	Number
Binaries	2456
em	6220
OB	16363
OBe	721
Non MK and uncertainty	16345

Note. These are not necessarily unique stars, as the same star can appear in different works, each reporting a (potentially) different spectral type

3.3 Methodology: Balanced Random Forest

In this section, we used the simplification of the encoded labels introduced in section 3.2.3 and the features detailed in section 4.2.2 to address the classification of massive stars as a supervised learning problem. We applied a Balanced Random Forest (BRF), an interpretable method (in terms of feature relevance, and that they provide a decision path based on observables, not highly non linear combinations of them) suitable for handling unbalanced training datasets (Chen et al., 2004). We trained

the *BalancedRandomForestClassifier* method from the *imbalanced-learn*[§] using the simplification of the encoded labels from Skiff within the OBA GDR3 sample and applied the classification to the unlabeled data.

3.3.1 Preprocessing Features

We removed sources with NaN values in any feature or duplicate J , H , and K_s photometry. The percentage of duplicate sources using the *tmass_oid* identifier was 0.013%. Then, we defined the minimum and maximum values for each feature by calculating the 1st and 99th percentiles of the encoded labeled data. We only selected the pool of new candidates that fell within these boundaries. This ensured that all unlabeled candidates were within the specified feature restrictions, as shown in Table .4.

We aimed to quantify how representative the OBA encoded labeled data are with respect to the pool of new candidates for each feature. In Fig. 3.2.4, we present a similarity test using the Wasserstein distance[¶] between these samples. In Fig. 3.2.5, we show the cumulative number of sources per bin of G magnitude.

The similarity test was normalized between the maximum and minimum distance for each bin of magnitude. The most distinct feature across all magnitude bins was AENS, followed by $\sigma_{\overline{w}}$, the second most important feature for the model (see Section 3.4). This feature reached a minimum at $G = 11$ mag and began to increase for both bright and faint stars. From Fig 3.2.5, we observed that for bright stars ($G = 10$ mag), the number of OBA encoded labeled sources at $G = 10$ mag is limited, which may explain the larger distance in this magnitude bin. For fainter stars, starting around $G \approx 12$, there was no noticeable increase in the cumulative number of OBA encoded labels. This suggested a bias in the labeled sample toward brighter stars, making extrapolation to fainter magnitudes uncertain. To be more conservative, we selected stars brighter than $G_{\text{mag}} \leq 12$ as a compromise between completeness and precision. The final sample consisted of 99388 stars up to $G = 12$ mag, of which 35465 had spectral types and 63923 were unlabeled.

3.3.2 Metrics

We evaluated the model's performance using Recall, Precision, F1-score, Geometric Mean, and Index Balanced Accuracy (IBA). In the text, we refer to completeness and purity as recall and precision, respectively. The last two metrics, designed for imbalanced datasets, were taken from the *imbalanced-learn* package (Lemaître et al., 2017)^{||}. The IBA incorporates any metric (M) and allows emphasis on

[§]BalancedRandomForestClassifier

[¶]https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html

^{||}<https://imbalanced-learn.org/stable/references/metrics.html>

specific classes through a weighting parameter (α). These metrics helped assess model construction, final classification, and comparison with other models. The metrics are presented as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.3.1a)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.3.1b)$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.3.1c)$$

$$\text{TNR} = \frac{TN}{TN + FP} \quad (3.3.1d)$$

$$\text{Dom} = \text{recall} - \text{TNR} \quad (3.3.1e)$$

$$\text{G-mean} = \sqrt{\text{Recall} \times \text{TNR}} \quad (3.3.1f)$$

$$\text{IBA}_\alpha(M) = (1 + \alpha \cdot \text{Dom}) \cdot M \quad (3.3.1g)$$

Where TP are the number of true positives, FP the false positives, FN the false negatives and TN true negatives. See subsections 3.3.3 and 3.4.1 for the use and results of these metrics.

3.3.3 Building the model

The random forest (RF) is a tree-based algorithm that consists of an ensemble of decision trees (Breiman, 2001). Each tree is constructed using a random subsample with replacement (a bootstrap sample) from the training set, and the final prediction is determined by a majority vote among all the trees.

The structure of a decision tree consists of a root node, internal nodes, and terminal (leaf) nodes. Data is classified based on decision rules inferred from features at different levels, from the root node (level 0) to the terminal nodes (maximum depth). A terminal node is reached when all training examples within it belong to the same class. These components are defined by hyperparameters: *n_estimators* indicates the number of decision trees, *max_features* denotes the subset of features used to determine the best split at each node, and *max_depth* sets the limit for recursive division.

Traditional RFs show high performance in different problems, however they underperform on highly imbalanced datasets since bootstrap samples often underrepresent the minority class. Instead, we used the Balanced Random Forest (BRF), a variation of RFs that modifies the sample distribution to ensure equal class representation in each tree. We selected the BRF from the *Imbalanced-learn* module, which offers various under-sampling strategies for bootstrapping: *majority*, *not minority*, *not*

Table 3.3.1: hyperparameters tested for the Random Forest and the best value

Features	Tested parameters	Best parameter
n_estimators	[500, 800, 1200]	1200
max_features	[sqrt, log2]	log2
max_depth	[10, 20, 30, None]	30
sampling_strategy	[majority, not majority, all]	majority

majority, *all*, and *auto*. We tested independent strategies for binary classification: 'majority', 'not majority', and 'all'.

To optimize BRF construction, we used Grid Search CV with stratified k-fold cross-validation. This method systematically explores hyperparameter combinations from a predefined grid. It divides the data into 'k' folds, using 'k-1' folds for training and the remaining fold for testing. We used a 5-fold stratified cross-validation and selected the IBA combined with $Gmean^2$ as the scoring metric, setting $\alpha = 0.1$, as recommended in García et al. (2012). Table 3.3.1 presents the tested and optimal hyperparameters.

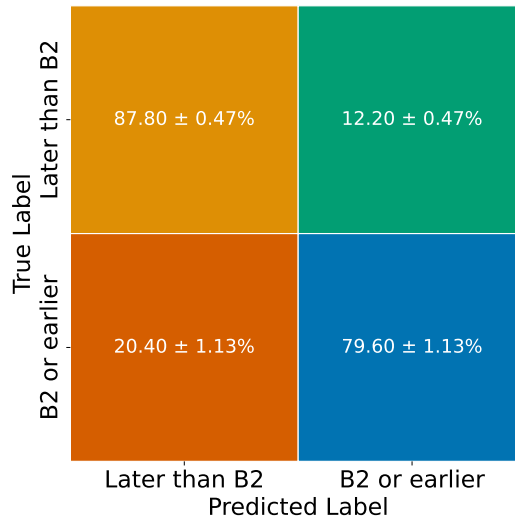


Figure 3.3.1: Confusion matrix for stars earlier than B2 and later than B2. The confusion matrix was obtained by generating 20 distinct training and testing sets, and training independent models on each. After training, the models were applied to their respective testing sets. We report the mean and standard deviation across all testing sets.

3.4 Results: Candidates from BRF

In the following, we use the built model and the preprocessed features to train a BRF and extend the classification to the unlabeled data. We also present the performance metrics of the algorithm and the moderate to high probability massive star candidates.

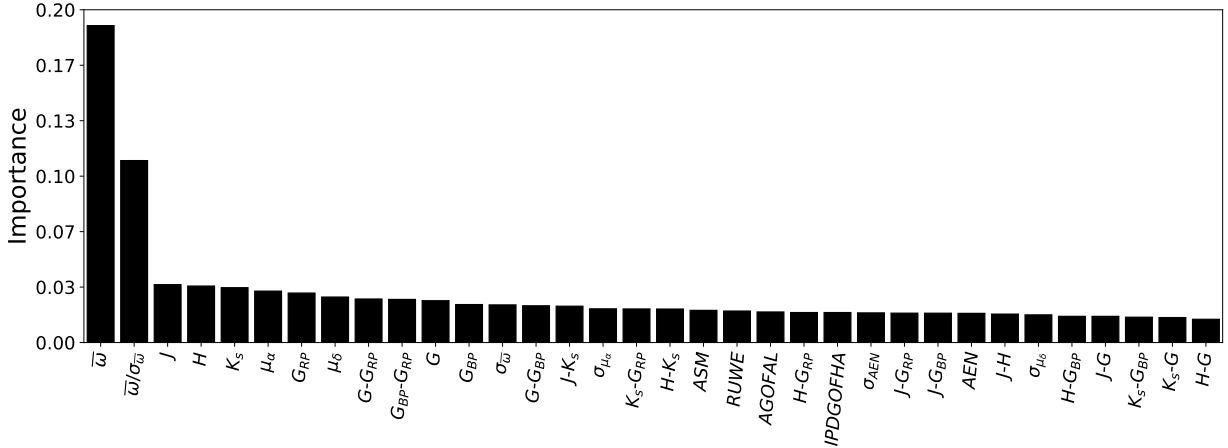


Figure 3.3.2: Feature importance for the BRF, sorted by significance. The importance is calculated as the mean and standard deviation of the accumulated reduction in variance across all trees.

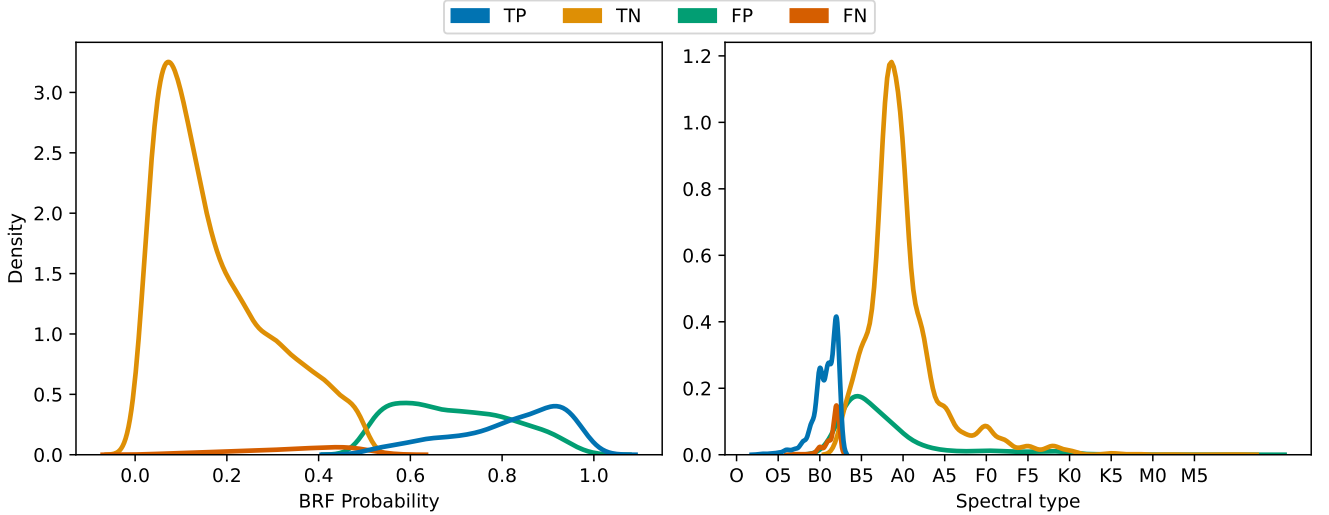


Figure 3.4.1: Kernel density plots for the metrics in the test set. The left panel shows the distribution in the probability of the BRF. The right panel shows the distribution of the different spectral types

3.4.1 Performance of the BRF

We defined the BRF using the best combination of hyperparameters and evaluated the performance of the model roughly following Sánchez-Sáez et al. (2021). We created 20 random permutations of the training and test sets using *StratifiedShuffleSplit* from *scikit-learn*, allocating 80% of the data for training and 20% for testing in each permutation. This corresponded to 28372 sources in the training set and 7093 in the test set per split. Across the full dataset, 4393 sources were classified as B2 or earlier, and 31072 as later than B2. Subsequently, we trained 20 BRFs, each on a separate splitting iteration using the training sample, and evaluated their performance with the respective test sets. We used a probability threshold of 0.6, above which stars are classified as spectral type B2 or earlier.

Figure 3.3.1 summarizes the results obtained for the 20 random permutations for each test set. TP,

TN, FP, and FN values are averaged over these permutations, with the error defined as their standard deviation. We obtained a TN fraction of $\sim 88\%$ and an FP fraction of $\sim 12\%$. The TP fraction was $\sim 80\%$, while the FN fraction was $\sim 20\%$. We measured a standard deviation of 0.47% for stars with spectral type later than B2, lower than that of stars with spectral type B2 or earlier.

Figure 3.3.2 shows the feature importance of the model, with parallax and parallax-over-error being the most significant. Infrared magnitudes from 2MASS follow in relevance, while Gaia magnitudes, colors, and proper motions rank next in importance. The goodness of fit of the astrometric solution has a minor contribution. Finally, one of the least informative features is the set of colors relating the G magnitude to infrared bands. As was expected, the parallax was the most important feature because, in combination with magnitudes, it serves to differentiate intrinsically bright objects from apparent ones.

The H–G color was the least informative for the model’s performance, although it traces the slope of the stellar spectrum. We explored this effect using the spectral energy distributions (SEDs) of 100 nearby sources (<1 kpc) with high predicted probabilities of being of spectral type B2 or earlier. These SEDs were fitted using VOSSA (?), allowing extinction to vary freely within the catalog provided limits. The results showed that H–G is highly degenerate with extinction, which explains its low relevance with respect to effective temperature.

Table 3.4.1: Performance metrics for the test sets

	Precision	Recall	F1 Score
Later than B2	0.968 ± 0.002	0.878 ± 0.005	0.921 ± 0.003
B2 or earlier	0.48 ± 0.01	0.80 ± 0.01	0.599 ± 0.009
Later than B3	0.948 ± 0.003	0.876 ± 0.005	0.910 ± 0.003
B3 or earlier	0.564 ± 0.009	0.77 ± 0.02	0.651 ± 0.009

Note. We include the B3 spectral type to illustrate the variation due to the differing definitions of a massive star.

We trained the final BRF classifier using a single data split, motivated by the small variation observed in the confusion matrices across different k -folds. All features were used to train the model, which was then applied to classify the OBA candidates in GDR3. We used 80% of the data for training and 20% for testing, applying the best hyperparameters from Table 3.3.1. Fig. 3.4.1 shows the kernel density estimates (KDE) for our test sets by metric. The left panel presents the BRF probability, while the right panel displays its distribution across spectral types. The TN peak was near 0.15, transitioning smoothly from lower probabilities before distorting as it approached 0.5, where the model had the highest classification confusion. For stars with probabilities above 0.6, TP and FP clearly overlapped. In the right panel, the metrics were associated with the spectral type, with the TN peak corresponding to A0 stars, while the FP peak was centered around B5 stars.

The peak in FP around spectral type B5 might be explained by the subtleties in spectral classification.

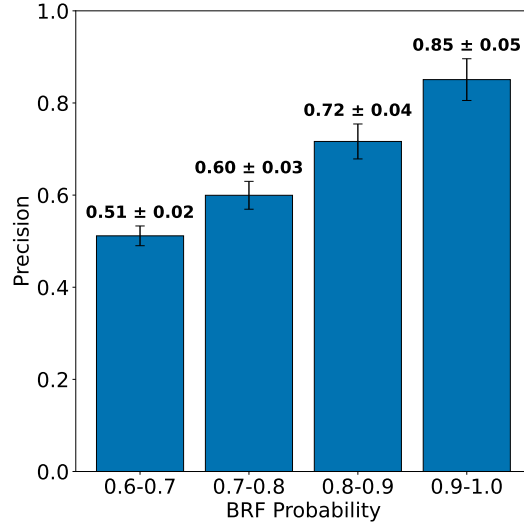


Figure 3.4.2: BRF precision for the test set in detecting B2 or earlier stars at different probability thresholds

The main spectroscopic difference between a B5 and a B2 star is the disappearance of silicon lines (very weak absorption features), accompanied by a strengthening of the Mg II 4481 line. Given that the model is informed only with photometry (“even worse”, broad band photometry), the model has not been exposed to data that traces such subtle differences and, therefore struggles to distinguish between these spectral types.

In addition, we performed the same analysis and did not find any obvious dependence on luminosity class in either the false positive or false negative cases.

Table 3.4.1 presents the performance metrics for the test sets. We observe high recall for stars with spectral types later than B2 and B2 or earlier. While precision remains high for stars later than B2, it is lower for B2 or earlier stars, indicating a high fraction of false positives. Fig. 3.4.2 shows how precision improves at higher probability thresholds. In addition, as noted in Section 3.2.3, the classification of massive stars based on spectral type is a working approximation, typically set at B2 or earlier. However, some authors, such as Zinnecker & Yorke (2007) (see Table 1), have considered B3 or early stars as massive. Therefore, we also computed the metrics under this criterion for comparison, observing an increase in precision. Finally, this trade-off between high completeness and lower purity can be addressed, as the final classification of massive stars requires spectroscopic confirmation.

3.4.2 Moderate to high probability massive stars candidates

From the sample of 99 388 stars with $G \leq 12$ mag, we classified the 63 923 unlabeled candidates using the BRF. We show the galactic distribution of our candidates in Fig. 3.4.3. The colors indicate the probability of classification as a star with spectral type B2 or earlier. We observed that the highest probabilities are more concentrated near the Galactic plane, even without RA and Dec coordinates as

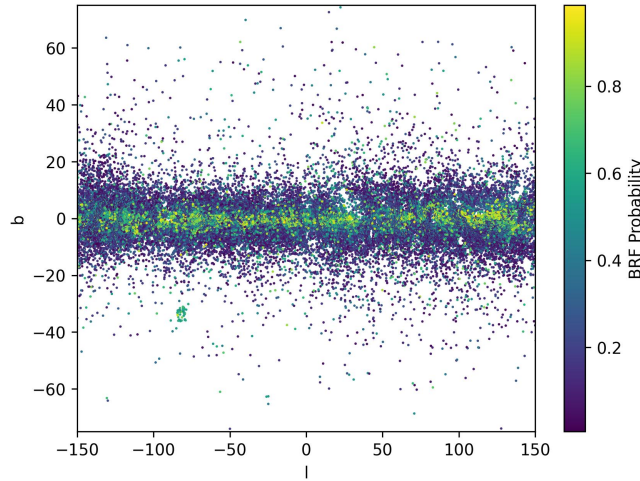


Figure 3.4.3: Galactic distribution of the candidate samples. This shows the distribution of stars outside our encoded labeled sample. The color represents the probability of being an star with spectral type B2 or earlier.

input for the algorithm.

Figure 3.4.4 shows the probability distribution for Em, OB, and OBe stars, which are the most abundant simplified encoded labels that were not included in the training or test sets. The distributions for OB and OBe stars are skewed toward higher probabilities, reflecting their predominance as early-type stars. In contrast, the Em distribution shows two peaks at low and high probabilities, suggesting the presence of both early- and late-type stars with emission. These trends suggest that the simplification of the encoded labels captured the global phenomenon without an evident systematic pattern for phase-variable spectral types.

Other spectral classes not considered in the training were excluded due to the feature space cuts applied during sample construction. As a result, we do not have a statistically significant number of objects from these classes to evaluate model performance reliably. In addition, assessing the classifier’s performance on Be stars is difficult because the Gaia and 2MASS photometry, the labels, and the spectroscopic data are non simultaneous. Since emission lines are time dependent features, this prevents us from drawing reliable conclusions. However, we expect Be stars to show distinct patterns in the Gaia feature space (?).

From the ~ 100000 stars analyzed, we identified 9089 moderate to high probability massive star candidates (probability ≥ 0.6) that were not included in our encoded labeled dataset. According to Table 3.4.1, this threshold corresponds to an expected precision of ~ 0.5 for identifying massive stars. Based on the FP and TP rates shown in Fig. 3.4.1, we expect the FP to be centered around B5 stars. This sample is more than twice the number of massive stars used for the development of the model, significantly expanding the set of potential massive stars and representing a valuable target for spectroscopic follow-up.

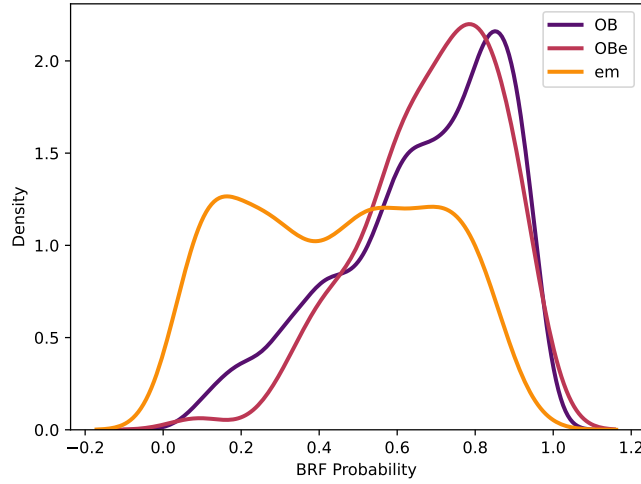


Figure 3.4.4: Kernel density plots for labeled data that are not included in our label codification.

3.5 Discussion

In the following we discuss the validation of our methodology incorporating external spectroscopic diagnostics, and the characteristics of our sample with respect to the literature.

3.5.1 Multiplicity considerations

The process of simplification of the numerical encoding of labels discussed in Section 3.2.3, implied the removal of multiple systems. However, as previously mentioned, multiplicity is a very common feature of massive stars (Offner et al., 2023). In order to assess the performance of our (singled source trained) algorithm in detecting multiple systems with a massive primary, we analysed the predictions of the model for a sample of 29 confirmed binaries from the Skiff catalogue. The number of systems was limited, as we only considered multiple systems that are independent of the model.

Figure 3.5.1 shows the spectral type of the primary stars against the BRF probability, with the spectral type of the secondary indicated by color. Systems with mass ratio ($q \sim 1$) have similar spectral types and are therefore expected to exhibit similar colors. We observe that most systems with a massive primary and $q \sim 1$ are identified by our methodology as massive indeed. On the other hand, the recovery efficiency is not 100% as two systems with a B primary and a close to 1 mass ratio lie just before our threshold probability. With this small test we can assess that, although we introduced a bias removing multiples from the training set, this does not seem to have a very strong effect (barring the small numbers of the experiment) in correctly predicting multiples with a massive primary.

3.5.2 Homogeneous validation dataset: LAMOST spectroscopy

We searched for spectral information in the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST, Zhao et al. 2012) to validate the classifier performance. This survey can obtain 4

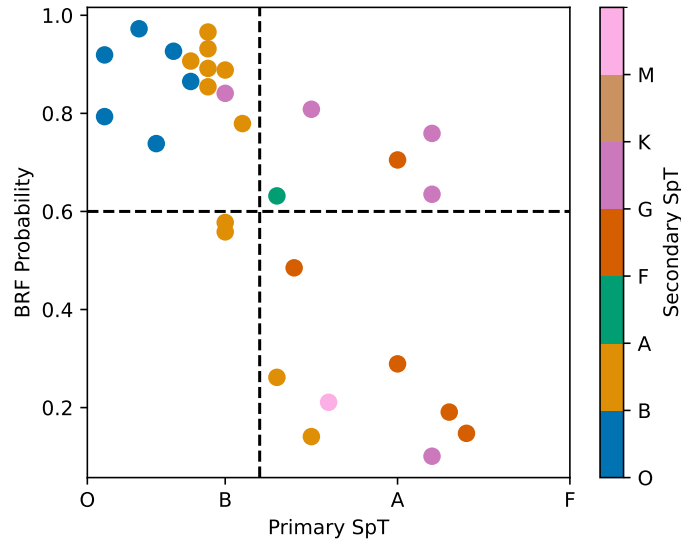


Figure 3.5.1: BRF probability of binary systems as a function of the primary’s spectral type. The spectral type of the secondary components is indicated by different colors.

000 spectra in one exposure for a limiting magnitude of 19 in *r* band and a spectral resolving power $R \sim 1800$ around the *g* band. It has a Field of View (FoV) larger than 20 square degrees and a fiber scale of 3.3 on the focal plane. The last public data release (DR10) ** contained low-resolution spectra for 11100139 stars, covering approximately $-10^\circ < \text{DEC} < 89^\circ$, making it a valuable database for large-scale spectroscopic studies.

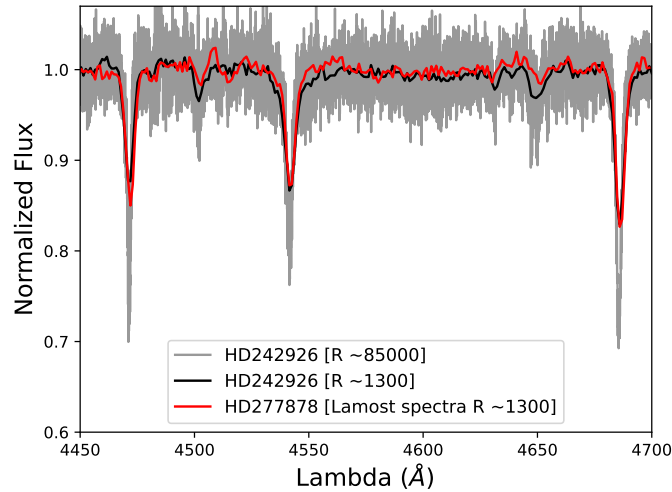


Figure 3.5.2: The massive standard star HD 242926, classified as O7Vz, is shown at high resolution (grayline) and degraded to LAMOST resolution (black line), in comparison to the O7V((f))z star HD 277878, observed by LAMOST (red line).

** <http://www.lamost.org/dr10/>

Automatic spectral classification

We found 1066 spectra in LAMOST within a 5radius for the moderate to high probability massive star candidates, considering all matches. To enhance the precision of our selection, we only considered counterparts located at the intersection of the initial 5radius and a 5wide band along the star’s proper motion direction. We discarded 106 spectra that showed anomalous offsets inconsistent with the expected proper motion direction. The final sample contained 880 spectra, from which we identified 585 unique sources. For each source with multiple observations, we retained the spectrum with the highest signal-to-noise ratio (SNR).

To derive spectral types for this sample, we developed an automatic methodology for classifying the low-resolution LAMOST spectra, based on a ground-truth sample of 50 massive standard stars of luminosity class V with spectral types O, B, and A collected from Didelon (1982); Sota et al. (2011); Maíz Apellániz et al. (2016); Martins (2018). The ground-truth spectra were obtained using the High-Efficiency and high-Resolution Mercator Echelle Spectrograph (HERMES; Raskin et al. 2011) and the Fiber-fed Extended Range Optical Spectrograph (FEROS; Kaufer et al. 1999), with resolutions of 85000 and 48000, respectively.

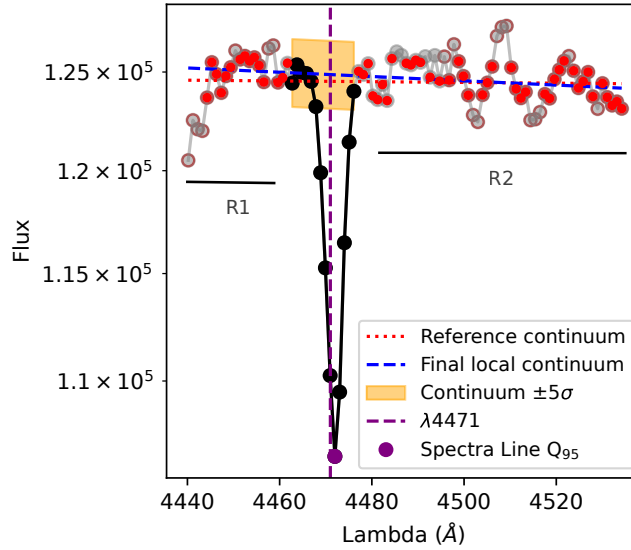


Figure 3.5.3: Automatic method to detect the continuum of the HeI $\lambda 4471$ spectral line. In gray, the points used to estimate the reference continuum. In red, full circles represent the points used to estimate the final local continuum. The Q_{95} (in purple) is used to determine if the peak or absorption exceeds 5σ above the continuum.

We degraded the high-resolution spectra of standard dwarfs to match LAMOST’s observational capabilities using the *specutils* package (Earl et al., 2024)^{††}. We applied convolution based smoothing to the spectrum using a Gaussian kernel, reducing the resolution to $R \sim 1300$, which is the expected resolution for LAMOST at HeI $\lambda 4471$ (see Figure 2 of Zhu et al. (2006)), an important line for dis-

^{††}<https://specutils.readthedocs.io/en/stable>

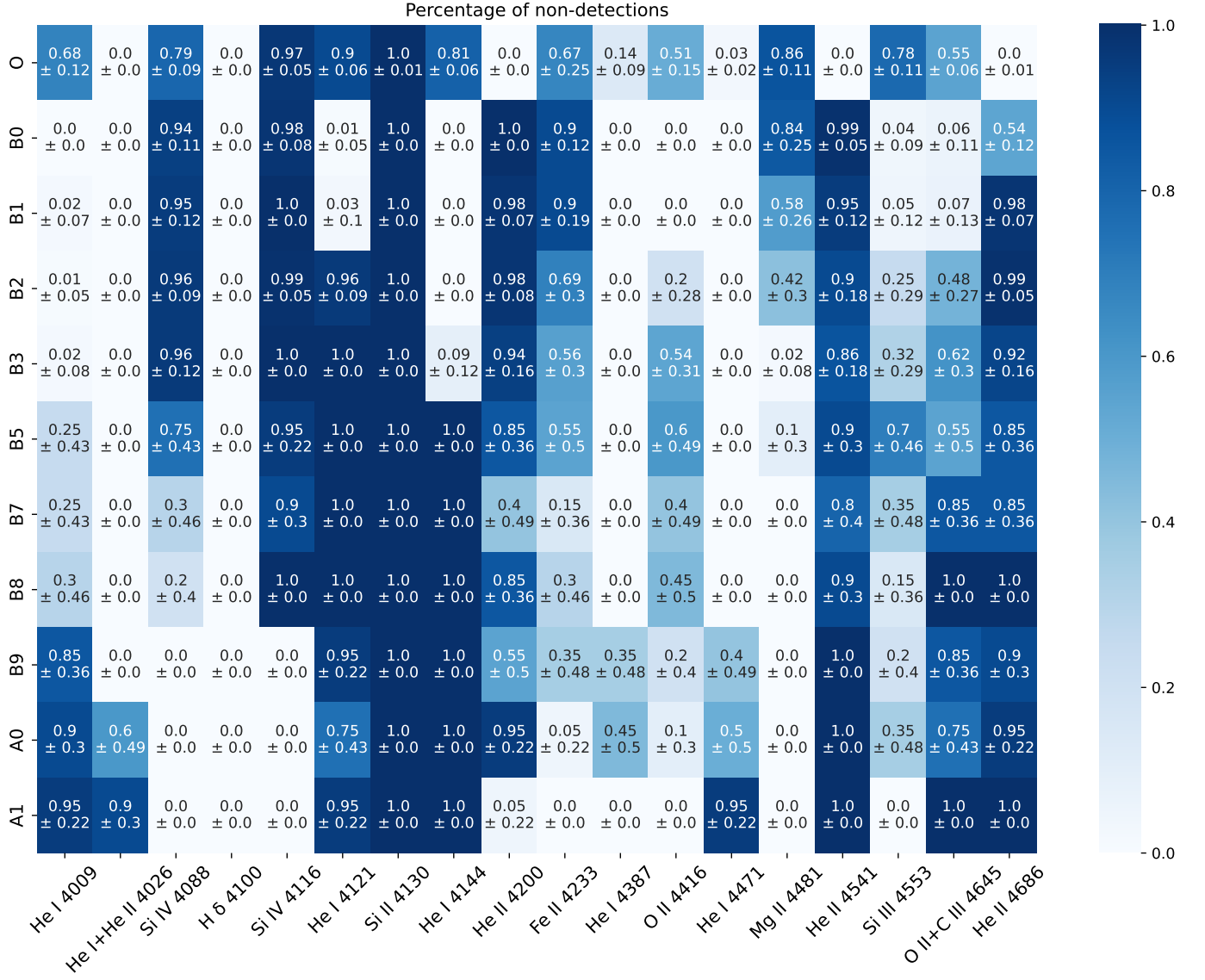


Figure 3.5.4: Percentage of non-detected lines across different spectral types of massive standard stars. For each iteration, we used a random sample of SNR values from the distribution of SNRs in LAMOST candidates. We present the mean detection rate and the standard deviation for 20 iterations.

tinguishing O and B stars. We resampled the flux to the LAMOST dispersion grid while preserving the integrated flux and added normal noise to compensate for the increase in SNR caused by the resolution reduction. We used a Wasserstein distance to compare the SNR distributions before and after adding noise. The distance decreased from 0.008 to 0.001, making the SNR of both distributions more similar. In Fig. 3.5.2, we show the spectrum of the massive standard star HD 242926, classified as O7Vz (Sota et al., 2011), as observed with HERMES, along with its degraded version at LAMOST resolution. For comparison, we also include the LAMOST spectrum of the O7V((f))z star HD 277878 (Roman-Lopes & Roman-Lopes, 2019).

We selected 18 spectral lines used in Kyritsis et al. (2022) and tested which lines were detected in the standard dwarfs with low resolution. For the detection of the lines, we combined the methodology presented in Bayo et al. (2011) and Kyritsis et al. (2022). The first approach detected the continuum of the line in a completely automatic way, while the second one used static spectral windows, taking into account the surroundings of each line.

We present an example of the detection of the HeI $\lambda 4471$ line in Fig. 3.5.3, the same procedure was applied to each line. We defined two fixed spectral regions (R1 and R2), in terms of central wavelength and width, to model the continuum, following Kyritsis et al. (2022), and a third one to trace the line (hereafter, the "line region"). An initial estimate of the continuum was derived from a linear fit to the R1 and R2 regions. This continuum was then refined by applying a one-sigma clipping. Once the final local continuum was identified, we assessed the significance of a detection by comparing the flux within the line region to the dispersion of the continuum. We subtracted the continuum and considered a detection whenever the peak of the line (defined as the 95 percentile of the flux distribution) was above 5 sigma.

We calculated the equivalent width (EW) of the line if it was detected. We considered that the line follows a Gaussian profile and calculated the associated σ . We performed three types of line fitting to characterize different scenarios: isolated line, blended line, and core emission line using a single Gaussian fit, a double Gaussian fit, and double inverted Gaussian fit, respectively. We selected the best fit using the Bayesian Information Criterion (BIC). This criterion is useful for model selection, giving priority to models with fewer parameters.

The line region was defined as $\pm 4\sigma$ around the center of the fitting model. We found that this threshold provided the best separation between the HeI $\lambda 4471$ and MgII $\lambda 4481$ lines in early-type stars when these lines are not blended. For consistency, the same criterion was applied to all other lines. Finally, we calculated the EW over the measured flux (not the fitted line), within the line region.

We applied the automatic methodology to characterize line detection in low-resolution spectra of dwarf massive standard stars. For a given spectral line, its detection can be influenced by the SNR of the spectra. To quantify the impact of SNR on line detection, we selected SNR values sampled from the distribution of moderate- to high-probability candidates and assessed the corresponding detections

in the low-resolution spectra of massive dwarf standard stars.

We performed 20 iteration, each using a different SNR value. In each experiment, we calculated the detection rate for all lines across the sample. The resulting matrix, shown in Fig. 3.5.4, reports the mean and standard deviation of the detection rates for each spectral type and line. This results are highly compatible with the results from Kyritsis et al. (2022). The Hei lines are clearly identified in B-type stars, while the Heii lines are systematically observed in O-type stars. The Si lines remain challenging to detect in both works.

From this matrix, we selected the following lines for classification: Hei $\lambda 4471$, Mgii $\lambda 4481$, Hei $\lambda 4144$, Heii $\lambda 4541$, Heii $\lambda 4686$, Feii $\lambda 4233$, Hei + Heii $\lambda 4026$. These lines are consistent with the expected spectral types, and based on them, we created an empirical decision path to classify the LAMOST spectra. It is important to note that we are only able to spectroscopically confirm stars of type B3 or earlier, due to the difficulty of detecting silicon lines at this resolution using an automatic approach, as illustrated in Fig. 3.5.4 and in agreement with Kyritsis et al. (2022).

Figure .1 summarizes the empirical decision path using automated line detection to classify O and early B stars. The first selection focused on Heii $\lambda 4686$ and Heii $\lambda 4541$, as these lines were consistently present in O stars. To minimize false positives due to the SNR in B-type stars, we applied a strict detection criterion for both lines. We then used the detection and EW ratio of Hei $\lambda 4471$ and Mgii $\lambda 4481$, which is useful to classify helium-normal stars later than B3 (Gray & Corbally, 2009), with a ratio near 1 for B6-B7 stars (Ramírez-Preciado et al., 2020). The Mgii $\lambda 4481$ line was consistently detected from B3 to A1, while Hei $\lambda 4471$ became undetectable from B9 onward and was almost always absent in A1 stars. We combined the detection of Hei + Heii $\lambda 4026$, which is always present in O and B stars, with the detection of Hei $\lambda 4144$, primarily observed up to B3. For stars later than A1, we used the Feii $\lambda 4233$ line, characteristic of A1 stars, and kept the designation "later" since we did not include spectra of standard stars of later types.

Estimated spectral types for the validation data-set

We applied the automatic methodology and empirical decision path to classify the LAMOST spectra of moderate to high probability massive star candidates. Fig. 3.5.5 presents the classification percentages, while Fig. 3.5.6 illustrates examples of spectra classified with an SNR in the g band detection in SDSS of approximately 250. We classified 57.2% of the sample as massive stars. Approximately 19.6% were misclassified as mid B-type stars (B4–B7). The results could improve by considering the luminosity class; however, developing an automated methodology for luminosity classification is beyond the scope of this work. We refer to Negueruela et al. (2024) for the standard classification of B stars. Additionally, 16% of the sample corresponds to stars later than type A, representing the lower limit of the model's misclassification.

Figure 3.5.7 illustrates the number of stars classified into different spectral type groups, separated by

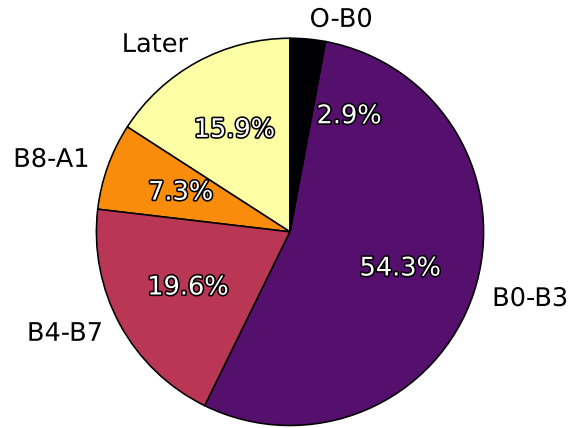


Figure 3.5.5: Spectroscopic classification percentages of moderate to high probability massive star candidates in the LAMOST sample, determined using the automatic methodology and empirical decision path.

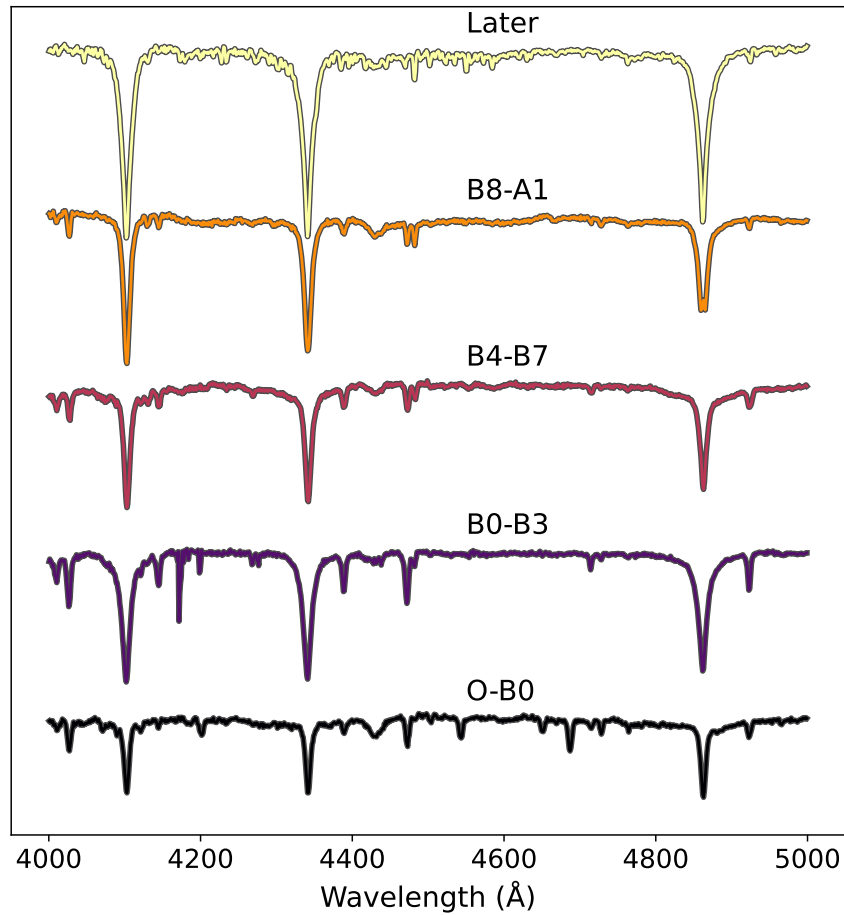


Figure 3.5.6: Examples of spectra from the LAMOST sample, classified into spectral type groups, with an SNR in the g band detection in SDSS of approximately 250.

the probability assigned by the BRF. At the top of each bar, we present the detection rate of massive stars, calculated as the percentage of the O-B0 and B0-B3 groups relative to all other spectral types within each probability interval. The detection rates align with the expected precision of 0.564 ± 0.009 for detecting B3 or earlier spectral types, as shown in Table 3.4.1. We observe that the detection rate increases with higher probabilities, while the total number of classified stars decreases. This reduction in moderate to high probability candidates is due to the selection function of the LAMOST survey, which does not cover the Galactic center, where a higher concentration of massive star candidates is expected.

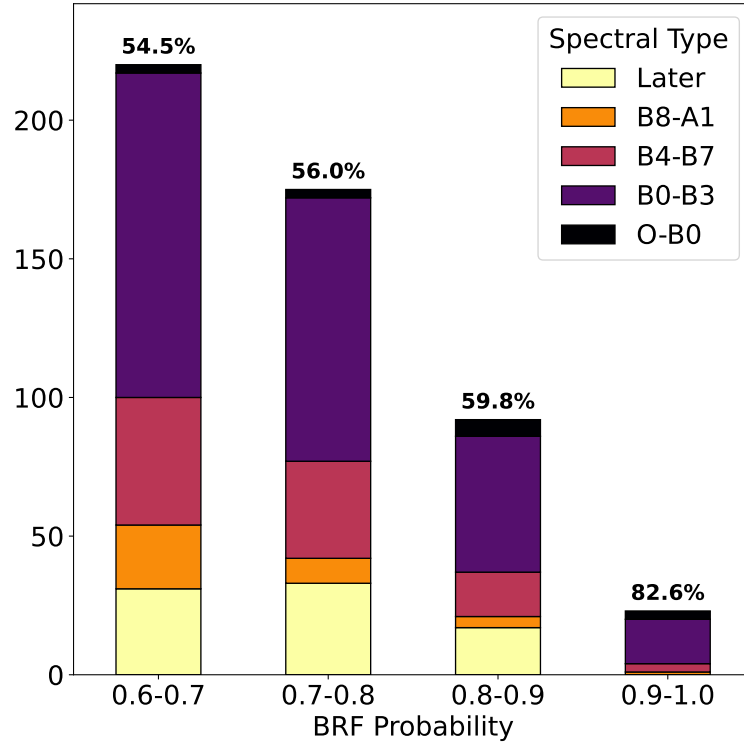


Figure 3.5.7: Number of stars classified in different spectral type groups, divided by the probability given by the BRF. The top of each bar shows the detection rate of massive stars.

3.5.3 Comparison with state-of-the-art massive star data bases and catalogs

We selected dedicated studies focused on the spectroscopic classification of massive stars to evaluate the performance of our BRF algorithm. Specifically, we examined the IACOB spectroscopic database (Simón-Díaz et al., 2011b), GOSSS (Maíz Apellániz et al., 2013), Chini et al. (2012), Li (2021), and the ALS compilation (Pantaleoni González et al., 2021). Within a 2search radius, we identified 124, 36, 17, 29, and 831 stars, respectively, that were not included in our training or test sets.

Figure 3.5.8 presents the probability distribution for different studies. The GOSSS and Li (2021) samples are more concentrated at high probabilities, as they consist exclusively of O-type stars. In the GOSSS sample, we classified all stars with a probability greater than 0.5, whereas in the Li (2021)

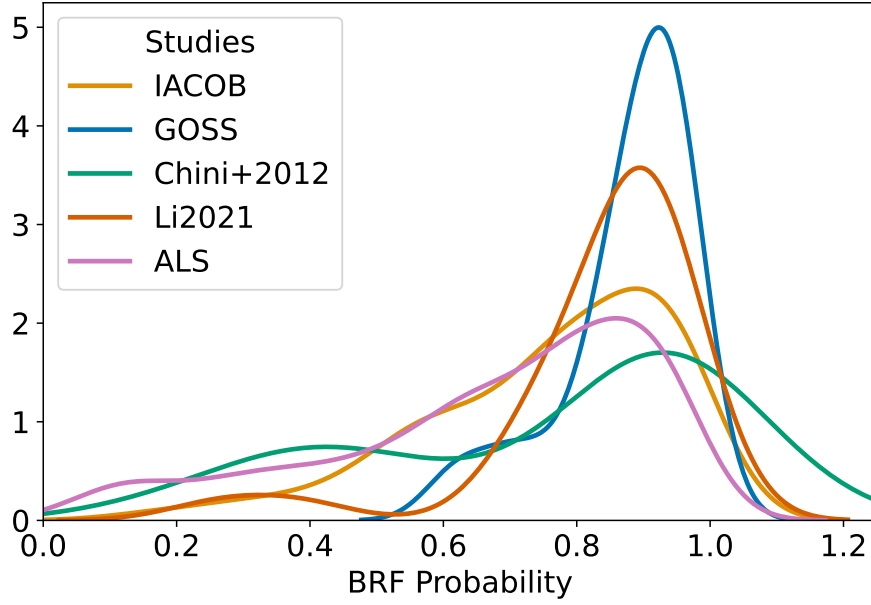


Figure 3.5.8: Probability distribution of the BRF for dedicated studies of massive stars.

sample, we misclassified two O9 stars. The comparison with the Chini et al. (2012), showed the poorest classifier performance, with most misclassifications resulting from intermediate B-type stars present in this sample. Among the five misclassified stars, three were B4 V stars, one was a B2 V star, and one was a B1 III star. In the IACOB and ALS samples, we classified approximately 10% and 20% of the stars, respectively, with probabilities below 0.5. However, not all stars in these studies had luminosity classifications, and some belonged to later B spectral types. In general, the probabilities were high, with a missed detection rate consistent with the algorithm’s expected metrics.

3.6 Summary and Conclusions

In this study, we present a moderate to high-probability catalog of massive star candidates with estimated spectral types B2 or earlier in the Milky Way. We defined our parental catalog using stars with spectral type O to A with astrometry and photometry from Gaia DR3, combined with photometry from 2MASS. We defined a subsample spectroscopically confirmed from the literature to serve as a source of labels. Since the labeled sample is biased toward brighter stars, we restricted our sample to $G < 12$ mag. In addition, we discarded possible outliers by discarding objects with any feature outside of the 1-99 percentile range.

A BRF classifier was trained on the encoded labeled data, simplifying the taxonomy to “B2 or earlier” or “later than B2” spectral types. The trained model was used to predict spectral types in this binary fashion on the unlabeled sample. The performance of our classifier shows high completeness of 0.8 and a purity ranging from 0.51 ± 0.02 for probabilities between 0.6 and 0.7, up to 0.85 ± 0.05 for the 0.9–1.0 range.

We mined spectroscopic archives for data on our moderate to high-probability candidates (estimated “B2 or earlier spectral type” with $P > 0.6$). The most homogeneous subsample obtained consisted of 534 objects with available spectra from the LAMOST survey. We developed an automatic methodology to detect the presence and characteristics of spectral lines on a sample of dwarf massive standard stars. Based on the extracted line(s) information, we designed a decision path capable of differentiating stars with spectral types O–B0, B0–B3, B4–B7, B8–A1, and later than A1. Applying this scheme to the 534 validation set, we obtained a confirmation rate of $\sim 54\%$. We further explored the dependence of this confirmation rate on the probability cut from the model, reaching 83% for $P > 0.9$. From the validation set (534 source), to the best of our knowledge 107 are completely new discoveries.

As a result of our work, we obtained a catalog of ~ 9000 moderate to high probability ($P > 0.6$) “B2 or earlier” bright ($G < 12$ mag) stars. From those, 5158 are not present in known massive stars databases (or general databases as Simbad). The remaining, 3931 were not present in our homogeneous set of encoded labels but were reported in other works in the literature. The number of new candidates presented in this work (5158) is therefore comparable to the currently known number of stars spectroscopically confirmed with spectral type B2 or earlier (7500 and 5793 for $G \leq 12$ mag). Assuming our metrics hold for the full sample, we expect to increase the known population of massive stars by approximately 50%.

Chapter 4

The variability of massive stars

This chapter presents the ongoing characterization of the variability of the sample proposed in Chapter 3. For this analysis, minor adaptations were performed to the methodology presented in Chapter 2. All results presented herein should be considered preliminary, as they require further in-depth analysis.

Abstract

The knowledge regarding internal structure of massive stars is more uncertain than that of intermediate and low-mass stars. This gap in understanding derives from the already mentioned scarcity of massive stars to be studied, but also from the historic lack of space-based photometric observations. This limitation was partially lifted by the advent of the Transiting Exoplanet Survey Satellite (TESS) that with its wide space coverage and dynamical range, it provides crucial datasets for investigating variability on timescales ranging from minutes to several tens of days. In this context, we present the results from an analysis of 3214 massive stars in the Galaxy with data from Sectors one to 72. Each sector was analysed individually, hence the period probed cover 0.01 to 12 days. Any periodic signal exceeding a false alarm probability threshold of 0.01 was phase-folded to that value and classified independently. This methodology produces sets of period-classification pairs (of different length) for each combination of object-sector. We combined these results by assigning the pair with the highest probability, or an unconstrained label when several pairs are statistically equally probable. The resulting variability distribution considering each sector-object observation as an independent member of a population is: 0.7% irregular variables, 15.1% eclipsing binaries, 27.1% ellipsoidal binaries, and 48% pulsating stars. When these results were grouped by object, we presented 621 eclipsing binaries, 1241 Ellipsoidal variables, 1348 pulsating stars, 195 unconstrained and 15 Non-periodic.

4.1 Introduction

Given the limitations in angular resolution that the typical distance to stars impose on our observational techniques, time resolved observations provide a unique opportunity to study stellar interiors and their evolution through the HR diagram (Catelan & Smith, 2015; Kurtz, 2022). Although the study of stellar variability is almost as old as the study of astrophysics itself, classically, it has been technically challenged with the limitations that ground based observations carry (aliasing, variable sensitivity following atmospheric changes, etc.). These limitations were largely mitigated with the advent of satellite missions targeted to acquire the most precise time series data in bulk as possible.

We have gained access to space-based photometry, with high stability, different time baselines and cadences, and nearly continuous light curves (LCs). For example, from the Convection, Rotation and planetary Transits (CoRoT; Baglin et al. 2009; Koch et al. 2010), the Kepler/K2 missions (Auvergne et al., 2009; Howell et al., 2014), and the Transiting Exoplanet Survey Satellite (TESS; Ricker et al. 2015). This has enabled the detection of all kinds of time-variable phenomena with precisions 10–100 times better than ground-based telescopes, including modes virtually undetectable from the latter (Kurtz, 2022).

High-precision space-based photometry has, therefore, revolutionised our ability to investigate stellar variability, with significant impact, for instance, in asteroseismology, the study of stellar interiors via surface pulsations (Aerts et al., 2010). In simple terms, asteroseismology studies compare the observed frequency spectra (extracted from the LCs) to those predicted by evolutionary models and thus infer the internal structures of the objects analyzed (Aerts, 2021). In this context, Kepler data have been very informative but biased towards low-mass stars given its main mission goal (transit detection) which determined sensitivity and sky coverage (neither suitable for the study of massive stars) (Deheuvels et al., 2012; Mosser et al., 2012; Yu et al., 2018). As a result, the mission provided only limited indirect observations of massive stars (Buysschaert et al., 2015; Aerts et al., 2017).

On the other hand, thanks to the TESS mission (whose primary goal is the detection of exoplanets), nearly the entire sky has become accessible, thereby opening new avenues for studying the variability of massive stars (Bowman, 2020, 2023). Various efforts have been undertaken to statistically identify large samples of variable early-type stars, including OBA and even F eclipsing binaries (IJspeert et al., 2021, 2024), and pulsating stars (Hey & Aerts, 2024). These results identify ~ 1000 eclipsing binaries and ~ 200 pulsating stars, based on literature-compiled O-type stars. In addition, studies aimed at identifying pulsations in spectroscopically confirmed binary systems (Southworth & Bowman, 2022; Eze & Handler, 2024) and eclipsing spectroscopic binaries (Pavlovski et al., 2023; Martín-Ravelo et al., 2024; Tkachenko et al., 2024) in samples of dozens of objects. These studies have provided, and continue to provide, crucial constraints for stellar structure and evolutionary models.

These observational samples help us to understand different open questions in the field of massive

stars, such as the mass discrepancy between observational mass estimates and model predictions (Herrero et al., 1992; Weidner & Vink, 2010). The blue supergiant problem, where there are more objects than expected from models (Bellinger et al., 2024; Kourniotis et al., 2025), and the convective core’s mass estimate, which strongly influences the evolution of massive stars (Burssens et al., 2023).

In this work, we present preliminary results on the distribution of different types of variability in a sample of ~ 3000 massive stars. We employed a Convolutional Neural Network model in combination with a Random Forest classifier. While the original workflow (CNN + RF) was trained on Optical Gravitational Lensing Experiment (OGLE), small modifications made it suitable to be applied to TESS data. Roughly, in the literature, covering similar spectral types, the variability has been characterized for ~ 1000 objects, even in the case of strong duplicity, our study has the potential to double the number of early B type stars characterized in terms of variability. This massive increase of population will allow us in the future, not only to have better statistics in the overall populations, but to identify outliers and exotic objects.

4.2 Sample, data and adaptation

4.2.1 OBA GDR3

We used the sample presented in Chapter 3 as input catalogue. This sample consists of massive star candidates (and previously known in the literature) selected, via a Balanced Random Forest (BRF, Chen et al. (2004)) on astrophotometric information from Gaia Data Release 3 (Gaia DR3; Gaia Collaboration et al. 2023), the Two Micron All Sky Survey (2MASS; Skrutskie et al. 2006). Previously reported sources were identified from the Skiff catalogue (Skiff, 2014), the IACOB spectroscopic database (Simón-Díaz et al., 2011b), GOSSS (Maíz Apellániz et al., 2013), and works by Chini et al. (2012), Li (2021), and the ALS compilation (Pantaleoni González et al., 2021).

4.2.2 The TESS data

The massive star input catalog includes 13299 sources, which we cross-matched with the TESS Input Catalog (TIC) V8 (Stassun et al., 2019), using a two arcsecond radius and retaining all possible counterparts. This yielded a total of 8224 stars, with 7716, 233, and 14 having one, two, and three counterparts, respectively. Based on the distance and Galactic distributions of the double matches, the vast majority are located in the densest regions, where some degree of crowding is expected. A flag was included in the cross-matched table to indicate the number of counterparts, and multiple TIC matches were retained. A total of 18 stars were associated with multiple TIC IDs while being resolved with different identifiers in the 2MASS Point Source Catalog. Thus, we can conclude that less than 3% of the multiple matches may be affected by contamination. In case of multiple 2MASS identifiers, given the astrometry and pixel size of this catalog, instead of propagating all counterparts,

we considered only the closest 2MASS object as the "right" counterpart. We note again that this choice should not have any significant impact in our results (as it affects less than 3% of the sources).

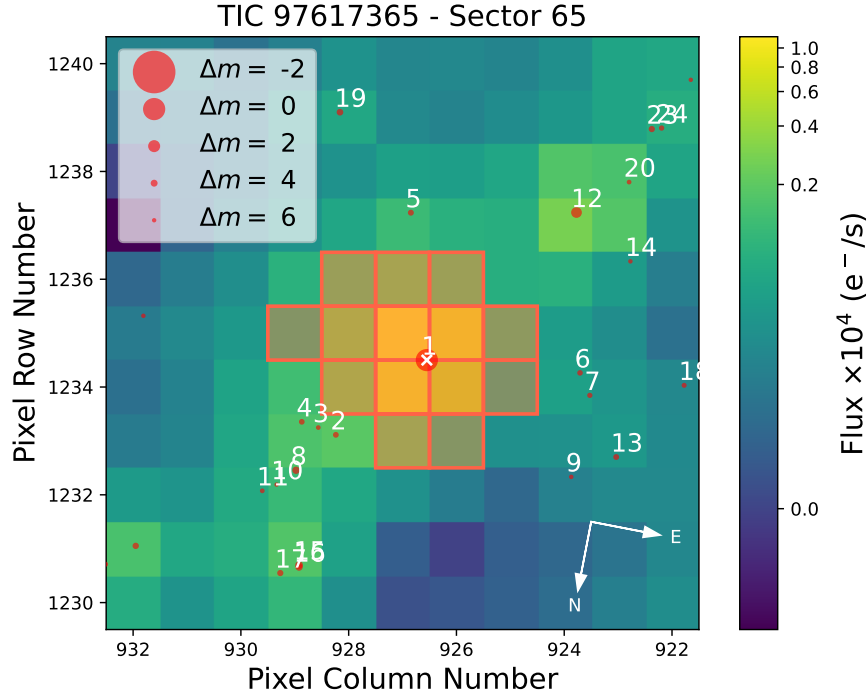


Figure 4.2.1: TPF plot from `tpfplotter` for HD 138729, an early B-type supergiant (Houk & Cowley, 1975). The orange outline shows the photometric aperture, and the numbers identify additional Gaia sources in the field of view with the symbol size representing the relative magnitude with respect to the TIC source.

We used Astroquery (Ginsburg et al., 2019) to download TESS LCs from the Mikulski Archive for Space Telescopes (MAST). In particular, we query by TIC identifier and access the products provided by the Science Processing operations Center (SPOC) pipeline. This pipeline provide calibrated target pixel files, simple photometry flux time series, .A total of 3769 stars had been observed in one or more sectors up to sector 72. We use each LC with its respective TIC and sector as an individual input, obtaining a final sample of 7896 star-sector entries.

The TESS mission have a large pixel scale (~ 21 arcseconds per pixel, Ricker et al. (2015)). Given this, we curated the LCs against contamination from neighbouring sources in the TESS aperture. To that effect, we used `tpfplotter` (Aller et al., 2020), which identifies the number of sources from the Gaia DR3 catalog in the TESS Target Pixel Files (TPFs). We first applied the default maximum magnitude of six from the target star to include possible contaminants in Gaia DR3 (see Fig. 4.2.1 as an example). Out of 3769 stars, 18 target stars were discarded because they did not fall within the aperture (although present in the input catalog, it does not have an associated LC, see Fig. A.3). For 1023 cases the science target was the only Gaia source within the aperture. The remaining cases, when more than one Gaia target fell within the aperture, we took the magnitude difference of the brightest target (with respect to the science target) and analyzed this different distribution searching

for an optimal threshold. As illustrated in Fig. 4.2.2, we adopted a difference of -2.5 mag as a balance between avoiding discarding too many sources but also only keeping contaminated cases in which at maximum the contaminant contributes with 10% of the measured flux. The final sample consist in 3420 stars.

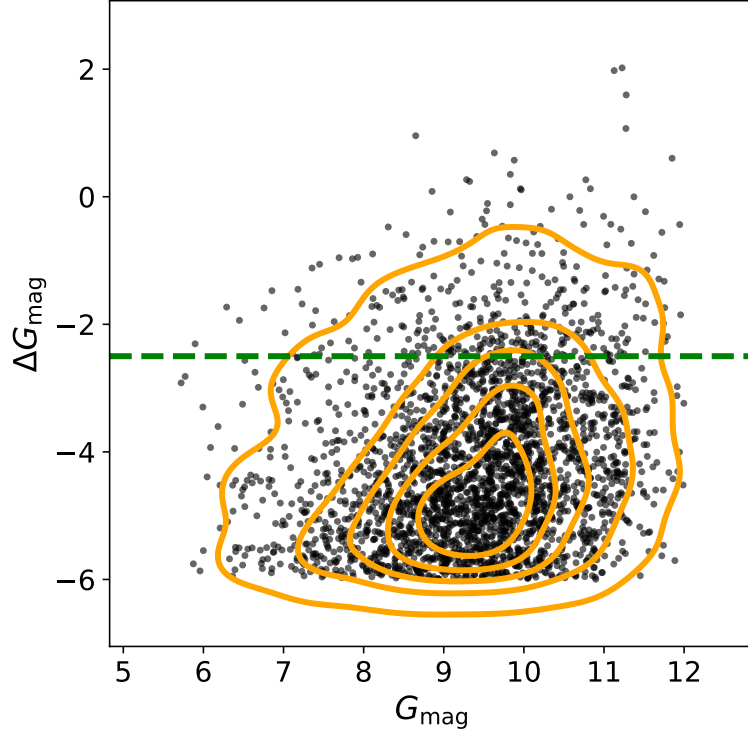


Figure 4.2.2: Difference in magnitude between the target star and the brightest object in the aperture against the Gaia G magnitude of the target. Gray dots represent the target stars; orange lines indicate density contours; and green dashed lines mark the threshold used to exclude contaminant stars.

4.2.3 Adaptation tests and searching for periodicity

We employed the combination of Convolutional Neural Network (CNN) and Random Forest (RF) classifier presented in Chapter 2 to classify the TESS LCs. In particular, from the combinations explored in Chapter 2, we adopted the CNN balanced by undersampling model (see Table 2.3.2).

In order to assess the suitability of this model "off-the-shelf" to a different kind of data, we studied its performance on two populations of labeled data: OGLE and The International Variable Star Index (VSX, Watson et al. 2006) stars with TESS observations, respectively. In addition to validating the transferability of the CNN model part of the methodology, we also modified the latter to probe possible periods taking advantage of the fast performance of the CNN (see below).

4.2.4 Period extraction

In order to classify the different types of variability in the sample of massive stars, we first determined the period of variation to construct the phase-folded light curves. We used a Lomb-Scargle periodogram (LS, Lomb (1976), Scargle (1982)) to calculate the periods. The minimum period was defined as twice the cadence, and the maximum period as half the observational baseline, with steps of 0.0001. These period limits varied depending on the applied quality mask and the observational sector.

As the algorithm is highly efficient, generating 76 images per second per core and predicting ~ 3000 images per second (see Chapter 2), we were able to evaluate a large number of phase-folded images. One image was generated for each significant peak of the periodogram, defined as any peak exceeding the 99% false alarm probability (FAP)* threshold.

The phase folded images were classified with the CNN+RF approach. The output classes corresponding to pulsating stars (Cepheids, Delta Scuti, Miras, and RR Lyrae) were combined into a single pulsating category. We selected the most reliable period as the one with the highest CNN+RF probability. If two phase-folded images had the same probability, we chose the one with the highest power in the periodogram. We referred to the high level of agreement sample as the set of star-sector entries with periods differing by less than 0.001 days.

Testing the methodology transfer: OGLE benchmark

We cross-matched the training sample from Section 2.3.2 filtered by periods shorter than 12 days against the SPOC database. This cross-matched resulted in a very modest sample of 20 stars. Applying the model to the TESS data using the period reported by OGLE correctly identified 14 Cepheids and 6 δ Scuti stars as pulsating, demonstrating good classification performance. Although the size of the sample is extremely limited, this first validation test shows promise.

This positive outcome was expected for two main reasons. First, the faster cadence of TESS versus the longer time baselines of OGLE translate in similarly dense phase folded lightcurves for many types of variables and yielded light curves with an average of ~ 7000 observations. This number is comparable to the typical number of observations in OGLE (see Fig. 2.3.2). Second, the OGLE I-band filter is in the Cousins system (Udalski et al., 2015)[†] and the TESS photometric system is defined to match this bandpass[‡]. As a result, for the same object, the expected amplitude of the TESS LC should be comparable to that of OGLE. That being said, obviously, differences in sensitivity between TESS and ground-based telescopes are expected.

To assess the reliability of the period estimation, we performed a comparison test with the 20 stars

*<https://docs.astropy.org/en/stable/timeseries/lombscargle.html#id38>

[†]SVO FPS Carlos Rodrigo

[‡]<https://tess.mit.edu/public/tesstransients/pages/readme.html>

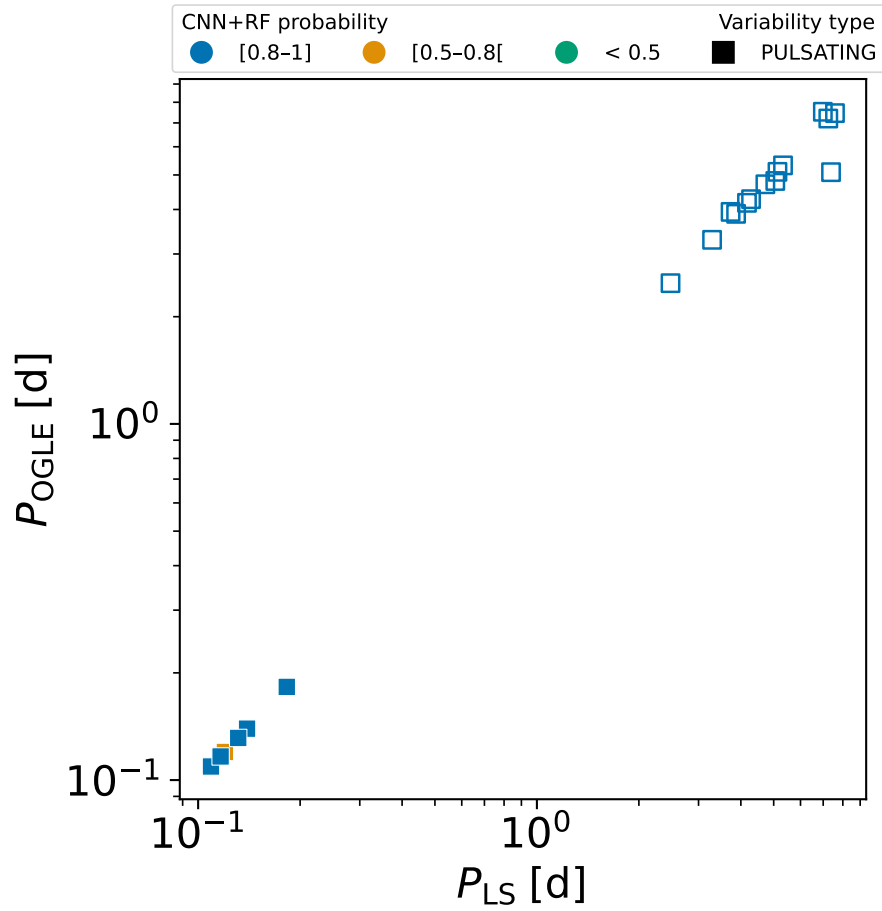


Figure 4.2.3: Comparison between the OGLE catalog period and the most probable period identified from the phase-folded images. Filled markers indicate the high level of agreement.

from the OGLE sample. From the LS we identified a total of 126 significant peak of the periodogram. Fig. 4.2.3 compares the OGLE period with the estimated realible period. The high level of agreement sample is presented with filled markers and are located along the diagonal line. We observe that, for all the stars, a similar period value is recovered and the high degree of agreement are concentrated to short period pulsation.

Testing the methodology transfer: VSX benchmark

We cross-matched the catalog of massive stars with available LCs against the VSX compilation, using a 2-arcsecond matching radius and retaining all counterparts, yielding a total of 1248 stars. Among these, 556 stars were reported classification and exhibited periods shorter than 12 days. From this subset, we retained only those stars for which the reported classification contained no uncertainty flags and indicated a single type of variability. The final sample comprises 307 stars, including 19 ellipsoidal variables, 109 eclipsing binaries, and 179 pulsating stars. Of the pulsators, 139 are classified as Beta Cepheids, while the remainder correspond to other reported types of pulsating stars. We obtained 815 star-sector entries, from the LS we identified a total of 14108 periodic signals, with each entry exhibiting at least one significant signal.

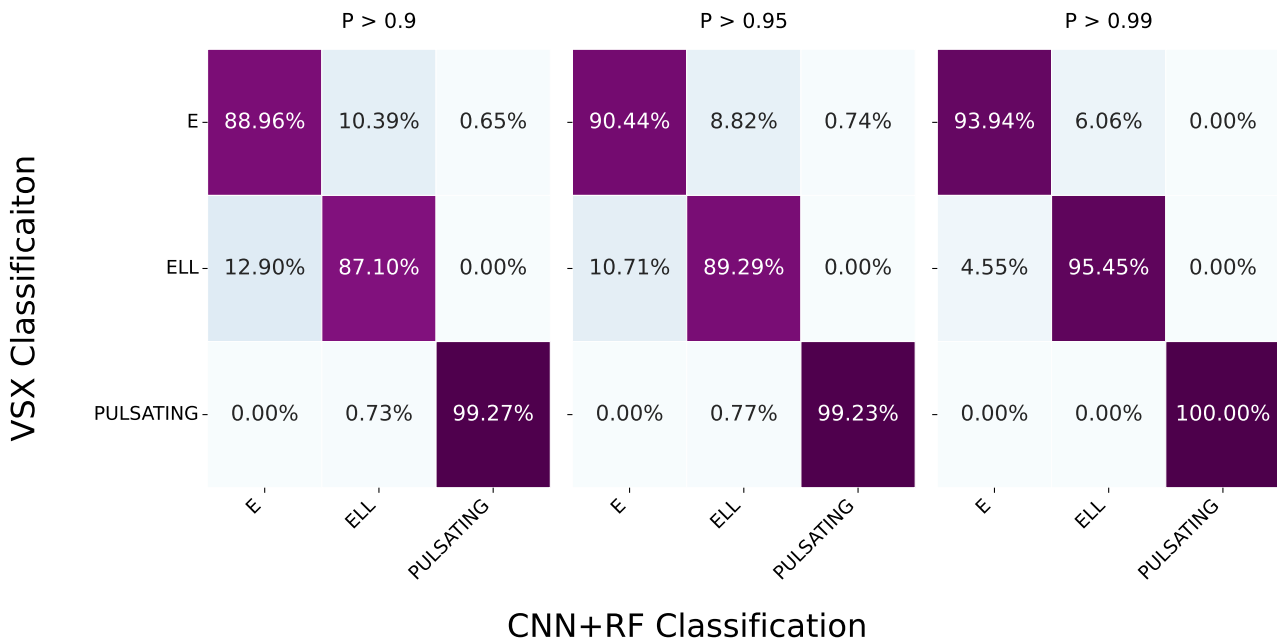


Figure 4.2.4: Confusion matrix for stars with variability classified by VSX compilation and predicted by the CNN+RF combination, evaluated at different probability thresholds.

Figure 4.2.4 presents the confusion matrix for different probability thresholds. The best performance is observed for pulsating stars, as the predicted probabilities for different subtypes are aggregated into a single category. Misclassification is observed between eclipsing binaries and ellipsoidal variables, as discussed in Chapter 2. In general, the results indicate a high level of performance, and the methodology is readily applicable to TESS LCs.

We compared the VSX compilation period with the estimated realible period to assess the agreement in the period identification, as shown in Fig. 4.2.5. In general, high-probability classifications [0.8–1] are located between two diagonal lines, suggesting the selection of harmonic periods, particularly for eclipsing binaries and ellipsoidal variables. Further analysis is be required to investigate this behavior. Nonetheless, there is overall good agreement between the VSX sample and the estimated periods.

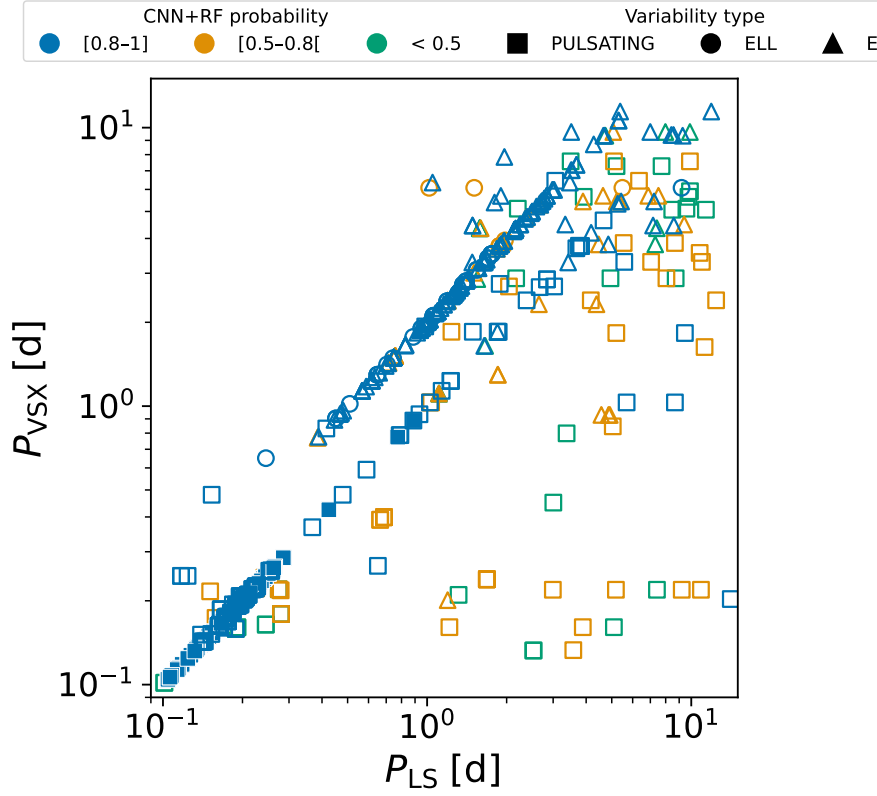


Figure 4.2.5: Comparison between the VSX catalog period and the most probable period identified from the phase-folded images. Filled markers indicate the high level of agreement.

4.3 Results

In this section, we presented the results of applying the classification pipeline, described in the previous sections, to all 7038 star-sector entries in the massive star sample. We identified 158082 possible periods and provided a periodic classification for 6312 star sectors.

Variability classes population

Figure 4.3.1 showed the percentage distribution of the different variability classifications. We defined non-periodic variability under two conditions: first, when no peak was detected above the FAP threshold; and second, when the probability of Random classes from the phase-folded images exceeded 0.8 across all sectors and peaks. We reported classifications for star-sectors with a probability greater than 0.8, while the remainder of the sample was reported as unconstrain.

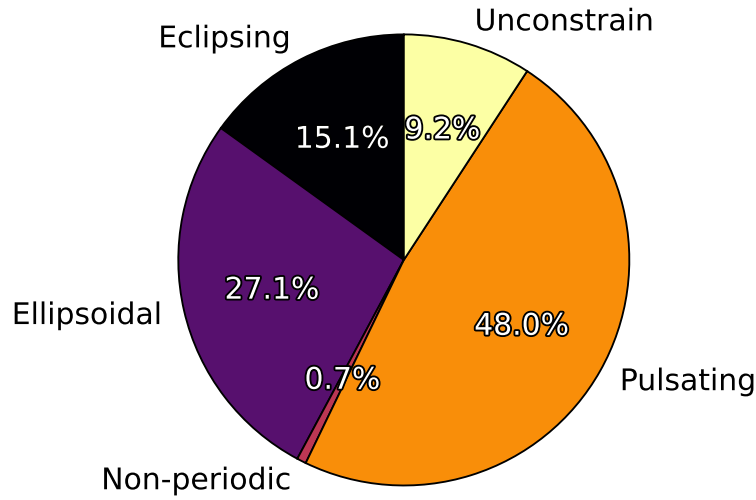


Figure 4.3.1: Variability distribution of the massive stars sample

We presented in Fig. 4.3.2 examples of classified pulsating stars, eclipsing binaries, and ellipsoidal variables. The model probability for the classification + period is extremely high in all cases (~ 1.0) and none of these targets have entries in the VSX compilation. The first column shows the light curve with the mean flux subtracted; the second shows the LS periodogram, with the FAP threshold and selected period indicated by red dashed lines, and the analyzed peaks marked in blue. The third column shows the phase-folded light curves using the selected period. In total, 65% of the objects with well-constrained classes and periods (according to our model) were not part of VSX, which translates into 2120 new constraints on variability for massive stars (candidates).

For stars classified in multiple sectors, we examined potential changes in their period. Fig. 4.3.3 shows the distribution of the standard deviation of the period, normalized by the mean period, for these multi-sector sources. We selected only those stars that received the same classification in all observed sectors. This is because ellipsoidal variables, which exhibit variability at half the orbital period, can be confused with pulsating stars whose detected period is approximately twice the fundamental pulsation period. Further analysis will be needed in these cases of harmonic misclassification. Stars in the tail of the distribution are potential candidates for binary interactions, evolutionary effects, or single stars that have undergone changes in periodicity.

4.4 Discussion

The results and methodology presented here demonstrate a preliminary exploitation of machine learning methods applied to massive stars. In this early stage, we present a sample that doubles the number of massive stars with characterized variability, even in cases of significant duplicity. In this direction of investigation, different improvements can be done.

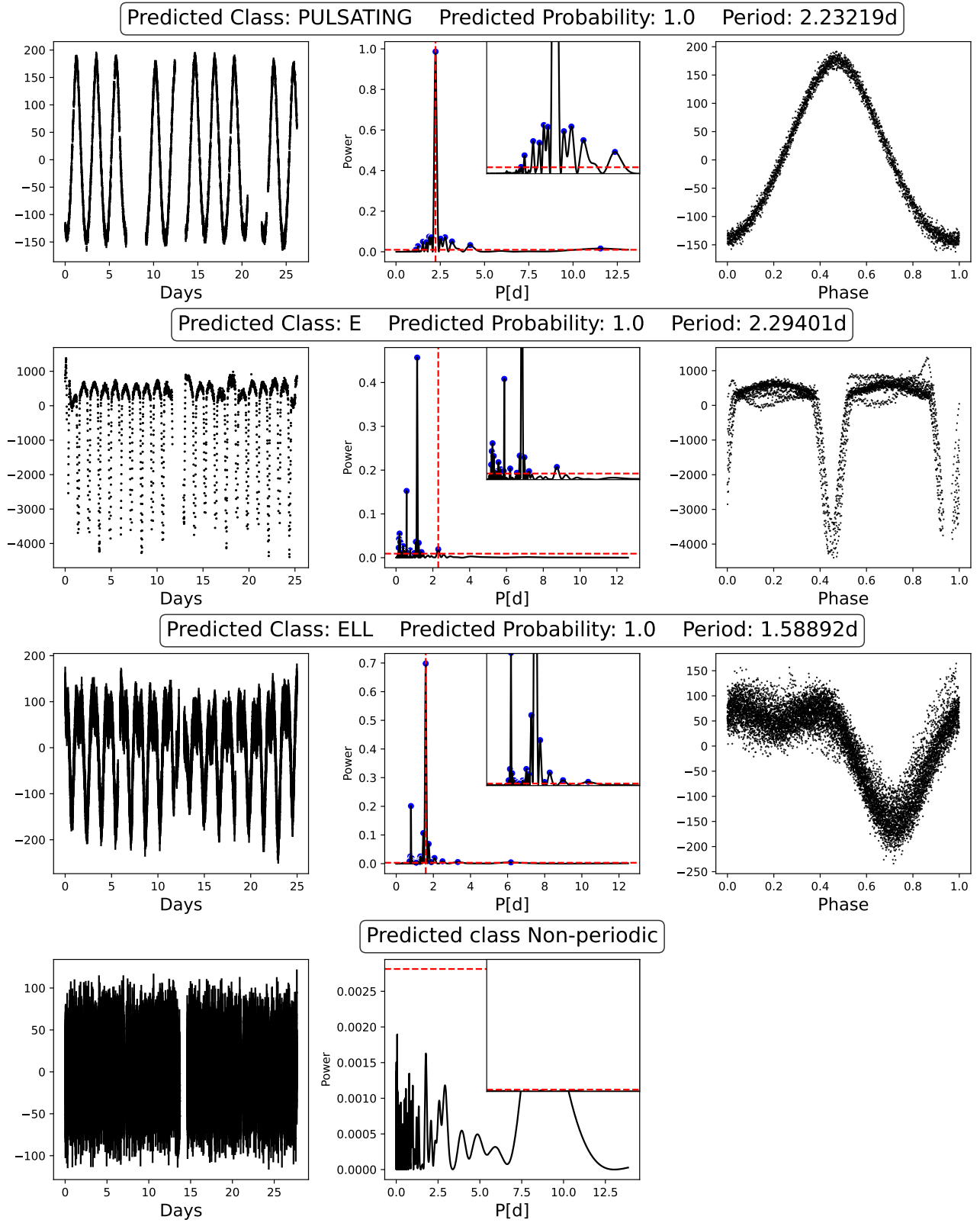


Figure 4.3.2: Examples of classified periodic variability. The first column shows the LC, the second displays the corresponding LS periodogram, and the third presents the phase-folded LC using the most reliable period

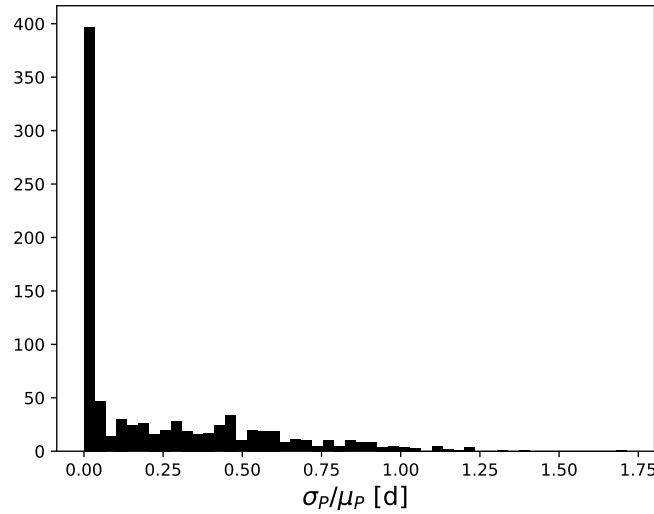


Figure 4.3.3: Distribution of the period dispersion for sources observed in multiple sectors.

4.4.1 Methodological Refinements

In this work, we analyzed only approximately 10% of the total available sample of massive stars described in Chapter 3. The number of analyzed stars can be significantly increased by incorporating additional publicly available bulk FFI LCs, such as those provided by the MIT Quick-Look Pipeline[§]. Furthermore, future improvements may focus on optimizing the integration of information from both CNNs and RFs, particularly for the effective isolation of harmonic signals. In addition, a methodology could be developed to combine LCs from different sectors, allowing for longer time baselines and the identification of variables with longer periods.

CNN as a periodogram filter

We presented, for the first time, the application of the CNN described in Chapter 2 as an accurate and efficient methodology for inclusion in periodogram analysis, which opens a new line of investigation. As a next step, we will explore the combination of different period-search algorithms, such as the Lafler-Kinman statistic Lafler & Kinman 1965, phase dispersion minimization Stellingwerf 1978; Scargle 1982, string length Dworetzky 1983, and analysis of variance methods Schwarzenberg-Czerny 1989, 1996. We aim to improve the detection rate and analyze its dependence on the pulsational and orbital periods, taking into account the pulsation mode and the type of binary system. This analysis will support an unbiased interpretation of the distribution of variability classes and their connection with the corresponding evolutionary states.

[§]DOI: 10.17909/t9-r086-e880

4.5 Summary and Conclusions

We have characterized the variability of 3420 objects, we have enlarged therefore the knowledge of massive star (candidate) variability by a factor of two in the worst case scenario. We propose a list of 621 new binaries whose orbits will shed light into important open questions in formation and evolution of massive stars. We have identified 195 objects with unconstrained class that deserve further study and objects with significant changes in period within months that also deserve further study.

The efficiency of the method proposed in Chapter 2 has been easy to adapt to this dataset with the advantage of probing a very large number of possible periods and discard harmonics and artifacts in a fast way

Chapter 5

Summary and conclusion

The aim of this thesis is to advance our astrophysical understanding of massive stars through the use of innovative machine learning approaches. Massive stars constitute a stellar population that has traditionally required careful, object-by-object analysis, due to their rarity and complexity. However, the vast amount of data now available in astronomy necessitates the development of efficient, scalable, and accurate solutions for the analysis of large, complex, and diverse datasets. By making use of these innovative technologies, we aim to identify and characterize a new population of massive stars. This contribution will serve as a catalyst for different branches of astronomy. From a Galactic perspective, it will improve predictions of the initial mass function and provide deeper insights into Galactic dynamics. From the stellar structure perspective, it will enhance the calibration of stellar evolution models and support the identification of pre- and post-interaction products, as well as contribute to our understanding of mass transfer processes. Finally, it will also aid in understanding the final stages of supernova progenitors and facilitate the study of exotic objects that may serve as sources of gravitational waves.

In Chapter 2, we proposed a methodology that transforms phase-folded light curves from the Optical Gravitational Lensing Experiment into images using two-dimensional histograms, which are then used for classification. The main objective was to automate the phase-folded visual inspection typically performed by traditional astronomers. We used a Convolutional Neural Network (CNN) to classify the 2D histograms of variable objects within eight different categories (extending to lower mass stars): Classical Cepheid, RR Lyrae, Long Period Variable, Miras, Ellipsoidal Binary, Delta Scuti, Eclipsing Binary, and spurious class with incorrect Periods (Rndm). We set up different training sets to train the same CNN architecture in order to characterize the impact of the training. Regarding computational performance, the image representation production rate is of ~ 76 images per core per second, and the time to predict is $\sim 60 \mu\text{s}$ per star. The accuracy of the classification improved from $\sim 92\%$, when based only on the CNN, to $\sim 98\%$ when the results of the CNN were combined with the period and amplitude features in a two step approach. This methodology achieves comparable results

with previous studies but with two main advantages: the identification of miscalculated periods and the improvement in computational time cost.

In Chapter 3, we presented a moderate to high-probability catalog of massive star candidates with estimated spectral types B2 or earlier. We defined our parental catalog using stars with spectral type O to A with astrometry and photometry from Gaia Data Release 3, combined with photometry from Two Micron All Sky Survey. We defined a subsample spectroscopically confirmed from the literature to serve as a source of labels. A BRF classifier was trained on the labeled data, simplifying the taxonomy to: “B2 or earlier” or “later than B2” spectral types. The trained model was used to predict spectral types in this binary fashion on the unlabeled sample. The performance of our classifier shows high completeness of 0.8 and a purity ranging from 0.51 ± 0.02 for probabilities between 0.6 and 0.7, up to 0.85 ± 0.05 for the 0.9–1.0 range. We mined spectroscopic archives for data on our moderate to high-probability candidates (estimated “B2 or earlier spectral type” with $P > 0.6$). The most homogeneous subsample obtained consisted of 534 objects with available spectra from the LAMOST survey. We developed an automatic methodology to detect the presence and characteristics of spectral lines on a sample of dwarf massive standard stars. Based on the extracted line(s) information, we designed a decision path capable of differentiating stars with spectral types O–B0, B0–B3, B4–B7, B8–A1, and later than A1. Applying this scheme to the 534 validation set, we obtained a confirmation rate of $\sim 54\%$. We further explored the dependence of this confirmation rate on the probability cut from the model, reaching 83% for $P > 0.9$. From the validation set (534 sources), to the best of our knowledge, 107 are completely new discoveries. As a result of our work, we obtained a catalog of ~ 9000 moderate to high probability ($P > 0.6$) “B2 or earlier” bright ($G < 12$ mag) stars. The number of new candidates presented in this work (5158) is therefore comparable to the currently known number of stars spectroscopically confirmed with spectral type B2 or earlier (7500 and 5793 for $G \leq 12$ mag). Assuming our metrics hold for the full sample, we expect to increase the known population of massive stars by approximately 50%.

In Chapter 4, we presented the preliminary results of the characterization of the variability of massive stars. We employed the methodology described in Chapter 2 to analyze a catalog of moderate- to high-probability massive star candidates with estimated spectral types of B2 or earlier. We characterized the variability of approximately 3000 massive star candidates, thereby doubling the number of massive stars with characterized variability in the most conservative scenario. We identified 621 eclipsing binaries, 1241 ellipsoidal variables, 1348 pulsating stars, and 195 unconstrained objects that warrant further investigation. The method introduced in Chapter 2 proved highly adaptable to this dataset, offering the advantage of efficiently probing a wide range of possible periods while rapidly eliminating harmonics and artifacts.

The present investigation increases the number of bright massive stars by approximately 50%, even in regimes where most such stars were previously considered to be identified. Furthermore, the preliminary results presented in Chapter 4 show that we have doubled the number of classified variable

massive stars, demonstrating that the application of machine learning techniques in this field significantly enhances our understanding of these rare objects.

There remains considerable potential for further development. From an astrophysical perspective, the methodologies presented here could be extended to include fainter stars. Additionally, integrating the predominantly optical data used in this study with information from other regions of the electromagnetic spectrum may lead to the discovery of particularly interesting objects, such as products of binary interactions.

This marks only the beginning of the large-scale characterization of massive stars. Most techniques employed in this work, such as Convolutional Neural Networks and Random Forests, are still considered traditional within the context of contemporary machine learning. Future work may incorporate transformer-based architectures, which have demonstrated excellent performance in other domains. These models could serve both to benchmark the traditional methods used here and to address broader astrophysical problems. We are now firmly in the era of big and complex data, which will bring numerous challenges, along with opportunities to uncover new and valuable astrophysical insights.

Chapter 6

Appendix

Table .1: Different works used in this paper for each variability class.

Variability Class	Reference
ELL	(Soszyński et al., 2016b)
M	(Soszyński et al., 2009b, 2011b) (Pietrukowicz et al., 2013b; Soszyński et al., 2013) (Iwanek et al., 2022)
CEP	(Soszyński et al., 2008, 2010a) (Soszyński et al., 2011c; Pietrukowicz et al., 2013b) (Soszyński et al., 2015; Udalski et al., 2018) (Soszyński et al., 2017, 2020)
DST	(Poleski et al., 2010; Pietrukowicz et al., 2013b) (Pietrukowicz et al., 2020; Soszyński et al., 2022) (Soszyński et al., 2023)
E	(Graczyk et al., 2011; Pietrukowicz et al., 2013a) (Pawlak et al., 2013; Soszyński et al., 2016b) (Pawlak et al., 2016)
LPV	(Soszyński et al., 2009b, 2011b) (Pietrukowicz et al., 2013b; Soszyński et al., 2013)
RR	(Soszyński et al., 2009a, 2010b) (Soszyński et al., 2011a; Pietrukowicz et al., 2013b) (Soszyński et al., 2014, 2016a) (Soszyński et al., 2019)

Table .2: Selected architecture and hyperparameters.

Layer Type	Layer Description
Input	Size: 32x32x1
Convolutional	No. of Filters: 16, Filter Size: 3, Activation: Relu, Padding: Same
Convolutional	No. of Filters: 16, Filter Size: 3, Activation: Relu, Padding: Same
Max-Pooling	Kernel Size: 2
Convolutional	No. of Filters: 32, Filter Size: 3, Activation: Relu, Padding: Same
Convolutional	No. of Filters: 32, Filter Size: 3, Activation: Relu, Padding: Same
Max-Pooling	Kernel Size: 2
Fully Connected	No. of Neurons: 1024, Activation: ReLu
Dropout	Dropout Probability: 30%
Fully Connected	No. of Neurons: 512, Activation: ReLu
Dropout	Dropout Probability: 30%
Fully Connected	No. of Neurons: 8, Activation: Softmax

Table .3: Categories of possible peculiarities:

Category	Peculiarity
Uncertainty	:, ?
Miscellaneous	comp, k, p, pec, !, var, v, w, wl, wk
Line-broadening	(n), n, nn, nnn, [n], s, sh, shell
Emission lines	e, em, (e), [e], eq, neb, ((f)), (f), f, ((f*)), (f*), f*, c, f?p, (n)fp, nfp, z, q
Carbon/Nitrogen enhanced	N, C

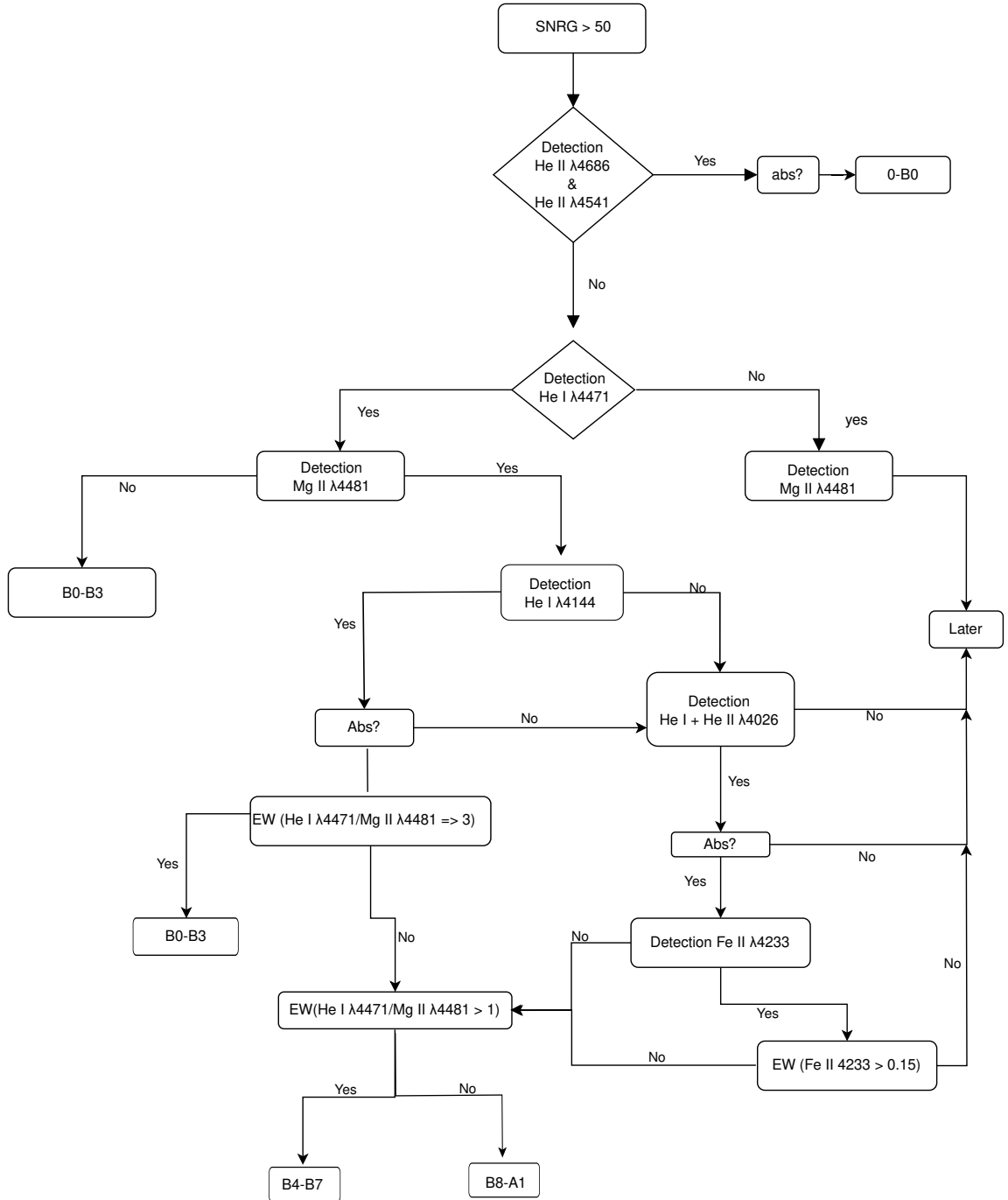


Figure .1: Empirical decision tree based on the massive standard stars. The decision is based on the detection and equivalent width of the He I, He II, Mg II, and Fe lines at low resolution ($R \sim 1300$).

Table .4: Minimum and maximum values for different features.

Min value	Max value	Feature
5.531	14.69	J
5.502	14.69	H
5.462	14.70	K_s
5.665	14.97	G
5.641	15.19	G_{BP}
5.611	14.73	G_{RP}
-0.125	0.433	$J-H$
-0.154	0.661	$J-K_s$
-2.426	0.3933	$J-G$
-3.246	0.4882	$J-G_{BP}$
-1.434	0.2395	$J-G_{RP}$
-0.077	0.261	$H-K_s$
-2.853	0.4913	$H-G$
-3.668	0.5923	$H-G_{BP}$
-1.856	0.3399	$H-G_{RP}$
-3.086	0.5268	K_s-G
-3.892	0.6251	K_s-G_{BP}
-2.087	0.3714	K_s-G_{RP}
-0.8725	0.2467	$G-G_{BP}$
-0.1592	1.046	$G-G_{RP}$
-0.2591	1.829	$G_{BP}-G_{RP}$
0.01088	0.4444	$\sigma_{\bar{\omega}}$
-2.518	161.8	$\bar{\omega}/\sigma_{\bar{\omega}}$
-0.1438	5.356	$\bar{\omega}$
-20.03	12.75	μ_{α}
-17.82	11.18	μ_{δ}
0.01095	0.4344	$\sigma_{\mu_{\delta}}$
0.00973	0.473	$\sigma_{\mu_{\alpha}}$
0.01693	0.7232	ASM
0.000	3.907	AEN
0.6365	20.98	$RUWE$
-9.306	265.8	$AGOFAL$
0.00265	0.1691	$IPDGOFHA$
0.000	1.564e+04	σ_{AEN}

Note. Minimum and maximum values for each features used in the analysis.

A Effects on the spectral type threshold in the BRF

The limit of massive stars can be different depending on the work and on the luminosity class, as we mention in Section 3.2.3. During the body of the work “B2” was used as a threshold. However, in this appendix we show the impact that a slight change in threshold to “B3” would have in our results. To this effect, we trained a new BRF following the same procedure described before but using B3 as the boundary between the two classes. The confusion matrix of this new model is shown in Fig. A.1. In addition, we also performed the same flow described in the text for external validation with LAMOST spectroscopy. The final results of this external validation are shown in Fig. A.2.

Even though the difference in performance of the algorithm are not striking when comparing the two confusion matrices (Figs 3.3.1 and A.1), the external validation procedure seems to indicate an underestimation of the contamination in the 0.6-0.7 probability bin for the case where B3 was used as a threshold. Therefore, thresholding with “B2”, provides more robust results.

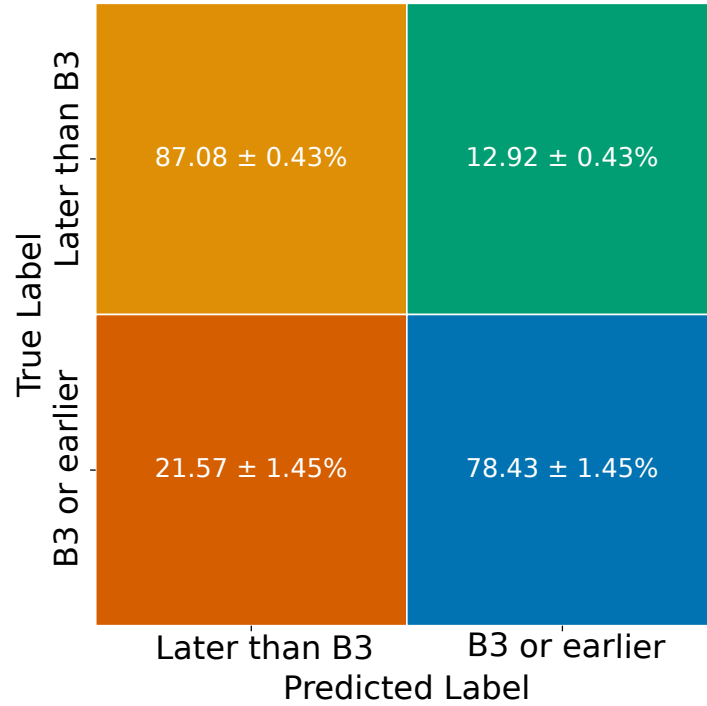


Figure A.1: Confusion matrix for stars earlier than B3 and later than B3. The confusion matrix was obtained by generating 20 distinct training and testing sets, and training independent models on each. After training, the models were applied to their respective testing sets. We report the mean and standard deviation across all testing sets.

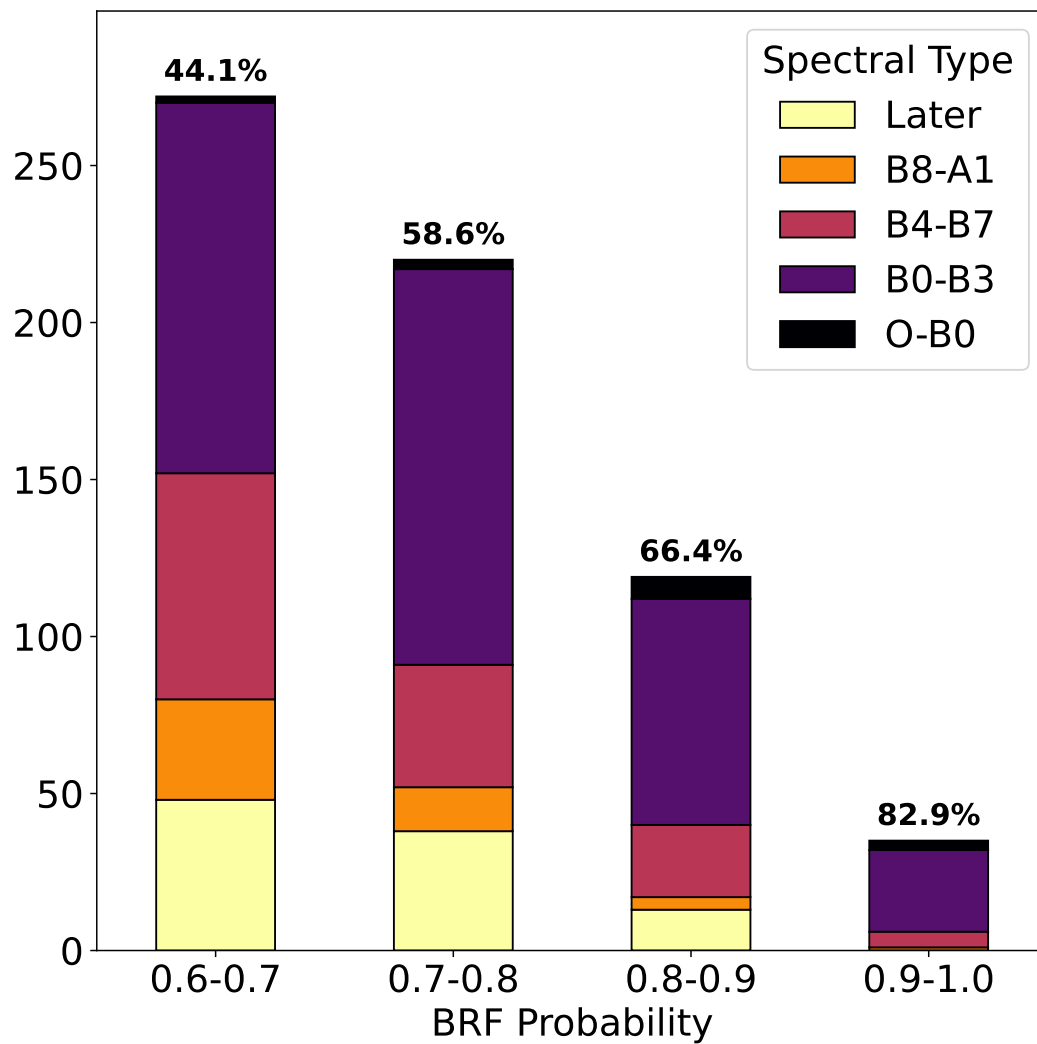


Figure A.2: Number of stars classified in different spectral type groups, divided by the probability given by the BRF using B3 thresholds candidates. The top of each bar shows the detection rate of massive stars.

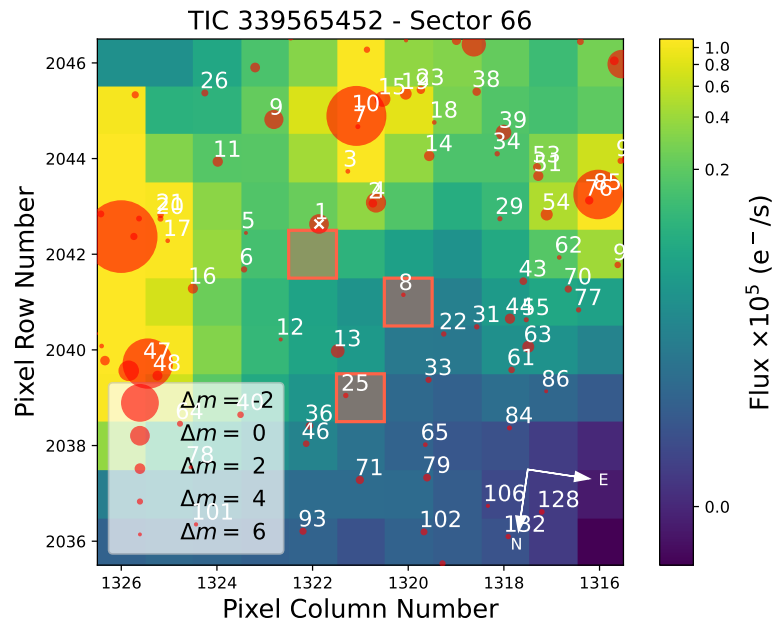


Figure A.3: Example of star listed in the input catalog that do not have an associated light curve

Bibliography

Aerts C., 2021, *Reviews of Modern Physics*, 93, 015001

Aerts C., Christensen-Dalsgaard J., Kurtz D. W., 2010, *Asteroseismology*, doi:10.1007/978-1-4020-5803-5.

Aerts C., et al., 2017, , 602, A32

Aguirre C., Pichara K., Becker I., 2019, , 482, 5078–5092

Alard C., Lupton R. H., 1998, , 503, 325–331

Alcock C., et al., 1997, , 486, 697–726

Aller A., Lillo-Box J., Jones D., Miranda L. F., Barceló Forteza S., 2020, , 635, A128

Auvergne M., et al., 2009, , 506, 411

Baglin A., 2003, *Advances in Space Research*, 31, 345–349

Baglin A., Auvergne M., Barge P., Deleuil M., Michel E., CoRoT Exoplanet Science Team 2009, in Pont F., Sasselov D., Holman M. J., eds, *IAU Symposium Vol. 253, Transiting Planets*. pp 71–81, doi:10.1017/S1743921308026252

Bailer-Jones C. A. L., Irwin M., Gilmore G., von Hippel T., 1997, , 292, 157–166

Bally J., Zinnecker H., 2005, , 129, 2281–2293

Barbá R. H., Gamen R., Arias J. I., Morrell N. I., 2017, in Eldridge J. J., Bray J. C., McClelland L. A. S., Xiao L., eds, *IAU Symposium Vol. 329, The Lives and Death-Throes of Massive Stars*. p. 89–96, doi:10.1017/S1743921317003258

Barbá R. H., Gamen R. C., Martín-Ravelo P., Arias J. I., Morrell N. I., 2022, , 516, 1149–1157

Bass G., Borne K., 2016, , 459, 3721–3737

Bastian N., Covey K. R., Meyer M. R., 2010, , 48, 339–389

Bayo A., et al., 2011, , 536, A63

- Becker I., Pichara K., Catelan M., Protopapas P., Aguirre C., Nikzat F., 2020, , 493, 2981–2995
- Bellinger E. P., de Mink S. E., van Rossem W. E., Justham S., 2024, , 967, L39
- Bellm E. C., et al., 2019, , 131, 018002
- Belokurov V., Evans N. W., Du Y. L., 2003, , 341, 1373–1384
- Belokurov V., et al., 2020, , 496, 1922–1940
- Berdyugina S. V., 2005, *Living Reviews in Solar Physics*, 2, 8
- Bergin E. A., Tafalla M., 2007, , 45, 339
- Bethe H. A., 1939, *Physical Review*, 55, 434
- Bethe H. A., Critchfield C. L., 1938, *Physical Review*, 54, 248
- Beuther H., Kuiper R., Tafalla M., 2025, arXiv e-prints, p. arXiv:2501.16866
- Blomme J., et al., 2010, , 713, L204–L207
- Bordé P., Rouan D., Léger A., 2003, , 405, 1137–1144
- Bowman D. M., 2020, *Frontiers in Astronomy and Space Sciences*, 7, 70
- Bowman D. M., 2023, , 368, 107
- Bowman D. M., et al., 2019, *Nature Astronomy*, 3, 760–765
- Breiman L., 2001, *Machine Learning*, 45, 5–32
- Brett D. R., West R. G., Wheatley P. J., 2004, , 353, 369–376
- Brott I., et al., 2011, , 530, A115
- Burssens S., et al., 2023, *Nature Astronomy*, 7, 913
- Buysschaert B., et al., 2015, , 453, 89
- Cannon A. J., Pickering E. C., 1901, *Annals of Harvard College Observatory*, 28, 129–P.6
- Catelan M., 2023, arXiv e-prints, p. arXiv:2302.01436
- Catelan M., Smith H. A., 2015, *Pulsating Stars*
- Cerviño M., Román-Zúñiga C., Luridiana V., Bayo A., Sánchez N., Pérez E., 2013a, , 553, A31
- Cerviño M., Román-Zúñiga C., Bayo A., Luridiana V., Sánchez N., Pérez E., 2013b, , 553, A32

- Chan H.-S., Villar V. A., Cheung S.-H., Ho S., O'Grady A. J. G., Drout M. R., Renzo M., 2022, , 932, 118
- Chandrasekhar S., 1931, , 74, 81
- Charbonneau D., Brown T. M., Latham D. W., Mayor M., 2000, , 529, L45–L48
- Chawla N. V., Japkowicz N., Kotcz A., 2004, SIGKDD Explor. Newsl., 6, 1–6
- Chen C., Liaw A., Breiman L., et al., 2004, University of California, Berkeley, 110, 24
- Chini R., Hoffmeister V. H., Nasserri A., Stahl O., Zinnecker H., 2012, , 424, 1925–1929
- Choudhuri A. R., 2010, Astrophysics for Physicists
- Cong S., Zhou Y., 2023, Artificial Intelligence Review, 56, 1905–1969
- Debusscher J., Sarro L. M., Aerts C., Cuypers J., Vandenbussche B., Garrido R., Solano E., 2007, , 475, 1159–1183
- Deheuvels S., et al., 2012, , 756, 19
- Didelon P., 1982, , 50, 199–207
- Ding X., Song Z., Wang C., Ji K., 2024, , 167, 192
- Dodd J. M., Oudmaijer R. D., Radley I. C., Vioque M., Frost A. J., 2024, , 527, 3076–3086
- Donoso-Oliva C., Becker I., Protopapas P., Cabrera-Vives G., Vishnu M., Vardhan H., 2023, , 670, A54
- Drake A. J., et al., 2014, , 213, 9
- Dubath P., et al., 2011, , 414, 2602–2617
- Dworetzky M. M., 1983, , 203, 917
- Earl N., et al., 2024, astropy/specutils: v1.19.0, doi:10.5281/zenodo.14042033, <https://doi.org/10.5281/zenodo.14042033>
- Eker Z., Bilir S., Soyduğan F., Gökçe E. Y., Soyduğan E., Tüysüz M., Şenyüz T., Demircan O., 2014, , 31, e024
- Eker Z., et al., 2015, , 149, 131
- Eldridge J. J., Stanway E. R., 2022, , 60, 455–494
- Eze C. I., Handler G., 2024, , 272, 25

- Farrell S. A., Murphy T., Lo K. K., 2015, , 813, 28
- Ferrario L., Pringle J. E., Tout C. A., Wickramasinghe D. T., 2009, , 400, L71–L74
- Folkes S. R., Lahav O., Maddox S. J., 1996, , 283, 651–665
- Förster F., et al., 2021, , 161, 242
- Gaia Collaboration et al., 2018, , 616, A1
- Gaia Collaboration et al., 2019, , 623, A110
- Gaia Collaboration et al., 2021, , 649, A6
- Gaia Collaboration et al., 2023, , 674, A1
- García V., Sánchez J., Mollineda R., 2012, *Knowledge-Based Systems*, 25, 13–21
- Ginsburg A., et al., 2019, , 157, 98
- Gonzalez O., et al., 2022, in Evans C. J., Bryant J. J., Motohara K., eds, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 12184, Ground-based and Airborne Instrumentation for Astronomy IX*. p. 1218412, doi:10.1117/12.2629932
- Graczyk D., et al., 2011, , 61, 103–122
- Graham M. J., Drake A. J., Djorgovski S. G., Mahabal A. A., Donalek C., Duan V., Maker A., 2013, , 434, 3423–3444
- Gray R. O., Corbally Christopher J., 2009, *Stellar Spectral Classification*
- Green M. J., Maoz D., Mazeh T., Faigler S., Shahaf S., Gomel R., El-Badry K., Rix H.-W., 2023, , 522, 29–55
- Götberg Y., de Mink S. E., Groh J. H., Leitherer C., Norman C., 2019, , 629, A134
- He K., Zhang X., Ren S., Sun J., 2016, in *Proceedings of the IEEE conference on computer vision and pattern recognition*. p. 770–778
- Heger A., Fryer C. L., Woosley S. E., Langer N., Hartmann D. H., 2003, *The Astrophysical Journal*, 591, 288
- Hellier C., 2001, *Cataclysmic Variable Stars*
- Hellings P., 1983, , 96, 37–54
- Henneco J., Schneider F. R. N., Hekker S., Aerts C., 2024, , 690, A65
- Herrero A., Kudritzki R. P., Vilchez J. M., Kunze D., Butler K., Haser S., 1992, , 261, 209

- Hey D., Aerts C., 2024, , 688, A93
- Houk N., Cowley A. P., 1975, University of Michigan Catalogue of two-dimensional spectral types for the HD stars. Volume I. Declinations -90° to -53°f0.
- Howell S. B., et al., 2014, , 126, 398
- Hubble E. P., 1925, Popular Astronomy, 33, 252–255
- Hubble E., 1929, Proceedings of the National Academy of Science, 15, 168–173
- IJspeert L. W., Tkachenko A., Johnston C., Garcia S., De Ridder J., Van Reeth T., Aerts C., 2021, , 652, A120
- IJspeert L. W., Tkachenko A., Johnston C., Aerts C., 2024, , 691, A242
- Ivezić , et al., 2019, , 873, 111
- Iwanek P., et al., 2022, , 260, 46
- Izzard R. G., Ramirez-Ruiz E., Tout C. A., 2004, Monthly Notices of the Royal Astronomical Society, 348, 1215–1228
- Jamal S., Bloom J. S., 2020, , 250, 30
- Jin S., et al., 2024, , 530, 2688–2730
- Kaufer A., Stahl O., Tubbesing S., Nørregaard P., Avila G., Francois P., Pasquini L., Pizzella A., 1999, The Messenger, 95, 8–12
- Kim D.-W., Bailer-Jones C. A. L., 2016, , 587, A18
- Kingma D. P., Ba J., 2014, arXiv e-prints, p. arXiv:1412.6980
- Koch D. G., et al., 2010, , 713, L79
- Kollmeier J., et al., 2019, in Bulletin of the American Astronomical Society. p. 274
- Kolmogorov A., 1933, Giorn. Ist. Ital. Attuari, 4, 1–11
- Kounkel M., et al., 2023, , 266, 10
- Kourniotis M., Cidale L. S., Kraus M., Ruiz Diaz M. A., Alberici Adam A., 2025, , 697, A152
- Kovács G., Zucker S., Mazeh T., 2002, , 391, 369–377
- Kroupa P., Weidner C., 2003, , 598, 1076–1078
- Kroupa P., Gjergo E., Jerabkova T., Yan Z., 2024, arXiv e-prints, p. arXiv:2410.07311

- Kurtz D. W., 2022, , 60, 31
- Kyritsis E., Maravelias G., Zezas A., Bonfini P., Kovlakas K., Reig P., 2022, , 657, A62
- Lafler J., Kinman T. D., 1965, , 11, 216
- Langer N., 2012, , 50, 107–164
- Lau M. Y. M., Hirai R., Mandel I., Tout C. A., 2024, , 966, L7
- Lea C., Vidal R., Reiter A., Hager G. D., 2016, in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part III* 14. p. 47–54
- Leavitt H. S., Pickering E. C., 1912, *Harvard College Observatory Circular*, 173, 1–3
- Lecun Y., Bengio Y., Hinton G., 2015, , 521, 436–444
- Lemaître G., Nogueira F., Aridas C. K., 2017, *Journal of Machine Learning Research*, 18, 1–5
- Li G.-W., 2021, *The Astrophysical Journal Supplement Series*, 253, 54
- Liu X.-Y., Wu J., Zhou Z.-H., 2009, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39, 539–550
- Lomb N. R., 1976, , 39, 447–462
- Mahabal A., Sheth K., Gieseke F., Pai A., Djorgovski S. G., Drake A. J., Graham M. J., 2017, in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, doi:10.1109/ssci.2017.8280984, <http://dx.doi.org/10.1109/SSCI.2017.8280984>
- Maíz Apellániz J., Negueruela I., Caballero J. A., 2024, *arXiv e-prints*, p. arXiv:2410.07301
- Marchant P., Bodensteiner J., 2024, , 62, 21–61
- Martins F., 2018, , 616, A135
- Martín-Ravelo P., Gamen R., Arias J. I., Chené A.-N., Barbá R. H., 2024, , 690, A306
- Martínez-Palomera J., Bloom J. S., Abrahams E. S., 2022, , 164, 263
- Matijević G., Prša A., Orosz J. A., Welsh W. F., Bloemen S., Barclay T., 2012, , 143, 123
- Maíz Apellániz J., et al., 2013, in *Massive Stars: From alpha to Omega*. p. 198 (arXiv:1306.6417), doi:10.48550/arXiv.1306.6417
- Maíz Apellániz J., et al., 2016, , 224, 4
- Minniti D., et al., 2010, , 15, 433–443

- Miroshnichenko A. S., et al., 2007, , 671, 828–841
- Mitchell T. M., Mitchell T. M., 1997, Machine learning. Vol. 1, McGraw-hill New York
- Moe M., Di Stefano R., 2017, , 230, 15
- Monsalves N., Jaque Arancibia M., Bayo A., Sánchez-Sáez P., Angeloni R., Damke G., Segura Van de Perre J., 2024, , 691, A106
- Morgan W. W., 1951, Publications of Michigan Observatory, 10, 33
- Morgan W. W., Keenan P. C., 1973, , 11, 29
- Morris S. L., 1985, , 295, 143–152
- Mosser B., et al., 2012, , 540, A143
- Muraveva T., Garofalo A., Scowcroft V., Clementini G., Freedman W. L., Madore B. F., Monson A. J., 2018, , 480, 4138–4153
- Nakar E., 2007, , 442, 166–236
- Naul B., Bloom J. S., Pérez F., van der Walt S., 2018, Nature Astronomy, 2, 151–155
- Negueruela I., Simón-Díaz S., de Burgos A., Casasbuenas A., Beck P. G., 2024, , 690, A176
- Nitish S., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R., 2014, Journal of Machine Learning Research, 15, 1929–1958
- Nun I., Protopapas P., Sim B., Zhu M., Dave R., Castro N., Pichara K., 2015, arXiv e-prints, p. arXiv:1506.00010
- Offner S. S. R., Moe M., Kratter K. M., Sadavoy S. I., Jensen E. L. N., Tobin J. J., 2023, in Inutsuka S., Aikawa Y., Muto T., Tomida K., Tamura M., eds, Astronomical Society of the Pacific Conference Series Vol. 534, Protostars and Planets VII. p. 275 (arXiv:2203.10066), doi:10.48550/arXiv.2203.10066
- Olmschenk G., et al., 2024, , 168, 83
- Packet W., 1981, , 102, 17–19
- Pantaleoni González M., Maíz Apellániz J., Barbá R. H., Reed B. C., 2021, , 504, 2968–2982
- Pantoja R., Catelan M., Pichara K., Protopapas P., 2022, , 517, 3660–3681
- Pavlovski K., Southworth J., Tkachenko A., Van Reeth T., Tamajo E., 2023, , 671, A139
- Pawlak M., et al., 2013, , 63, 323–338

- Pawlak M., et al., 2014, , 64, 293–308
- Pawlak M., et al., 2016, , 66, 421–432
- Pedregosa F., et al., 2011, *Journal of Machine Learning Research*, 12, 2825–2830
- Percy J. R., 2007, *Understanding Variable Stars*
- Pichara K., Protopapas P., Kim D. W., Marquette J. B., Tisserand P., 2012, , 427, 1284–1297
- Pickering E. C., 1890, *Annals of Harvard College Observatory*, 27, 1–388
- Pietrukowicz P., et al., 2013a, , 63, 115–133
- Pietrukowicz P., et al., 2013b, , 63, 379–404
- Pietrukowicz P., et al., 2020, , 70, 241–263
- Pojmanski G., 2002, , 52, 397–427
- Poleski R., et al., 2010, , 60, 1–16
- Prialnik D., 2009, *An Introduction to the Theory of Stellar Structure and Evolution*
- Ramírez-Preciado V. G., Roman-Lopes A., Román-Zúñiga C. G., Hernández J., García-Hernández D. A., Stassun K., Stringfellow G. S., Kim J. S., 2020, , 894, 5
- Raskin G., et al., 2011, , 526, A69
- Richards J. W., et al., 2011, , 733, 10
- Richards J. W., et al., 2012, , 744, 192
- Ricker G. R., et al., 2015, *Journal of Astronomical Telescopes, Instruments, and Systems*, 1, 014003
- Riess A. G., et al., 2011, , 730, 119
- Riess A. G., et al., 2016, , 826, 56
- Riess A. G., Casertano S., Yuan W., Macri L. M., Scolnic D., 2019, , 876, 85
- Rivinius T., Carciofi A. C., Martayan C., 2013, , 21, 69
- Roman-Lopes A., Roman-Lopes G. F., 2019, , 484, 5578–5586
- Rosen A. L., Offner S. S. R., Sadavoy S. I., Bhandare A., Vázquez-Semadeni E., Ginsburg A., 2020, , 216, 62
- Ruder S., 2016, arXiv e-prints, p. arXiv:1609.04747

- Rybizki J., Just A., 2015, , 447, 3880–3891
- Sana H., et al., 2012, *Science*, 337, 444
- Sanders J. L., 2023, , 523, 2369–2398
- Scargle J. D., 1982, , 263, 835–853
- Schneider P., 2006, *Extragalactic astronomy and cosmology: an introduction*. Vol. 146, Springer
- Schneider R., Ferrari V., Matarrese S., Portegies Zwart S. F., 2001, , 324, 797–810
- Schwarzenberg-Czerny A., 1989, , 241, 153
- Schwarzenberg-Czerny A., 1996, , 460, L107
- Shibayama T., et al., 2013, , 209, 5
- Shimizu R., Asako K., Ojima H., Morinaga S., Hamada M., Kuroda T., 2018, in 2018 First International Conference on Artificial Intelligence for Industries (AI4I). p. 27–30, doi:10.1109/AI4I.2018.8665709
- Simón-Díaz S., Garcia M., Herrero A., Maíz Apellániz J., Negueruela I., 2011a, in *Stellar Clusters & Associations: A RIA Workshop on Gaia*. p. 255–259 (arXiv:1109.2665), doi:10.48550/arXiv.1109.2665
- Simón-Díaz S., Castro N., Garcia M., Herrero A., Markova N., 2011b, *Bulletin de la Societe Royale des Sciences de Liege*, 80, 514–518
- Simón-Díaz S., et al., 2015, in Cenarro A. J., Figueras F., Hernández-Monteagudo C., Trujillo Bueno J., Valdivielso L., eds, *Highlights of Spanish Astrophysics VIII*. p. 576–581 (arXiv:1504.04257), doi:10.48550/arXiv.1504.04257
- Skiff B. A., 2014, *VizieR Online Data Catalog*, p. B/mk
- Skrutskie M. F., et al., 2006, , 131, 1163–1183
- Smartt S. J., 2009, , 47, 63–106
- Smirnov N., 1948, *The annals of mathematical statistics*, 19, 279–281
- Soszyński I., et al., 2008, , 58, 163–185
- Soszyński I., et al., 2009a, , 59, 1–18
- Soszyński I., et al., 2009b, , 59, 239–253
- Soszyński I., et al., 2010a, , 60, 17–39

- Soszyński I., Udalski A., Szymański M. K., Kubiak J., Pietrzyński G., Wyrzykowski , Ulaczyk K., Poleski R., 2010b, , 60, 165–178
- Soszyński I., et al., 2011a, , 61, 1–23
- Soszyński I., et al., 2011b, , 61, 217–230
- Soszyński I., et al., 2011c, , 61, 285–301
- Soszyński I., et al., 2013, , 63, 21–36
- Soszyński I., et al., 2014, , 64, 177–196
- Soszyński I., et al., 2015, , 65, 297–312
- Soszyński I., et al., 2016a, , 66, 131–147
- Soszyński I., et al., 2016b, , 66, 405–420
- Soszyński I., et al., 2017, , 67, 297–316
- Soszyński I., et al., 2019, , 69, 321–337
- Soszyński I., et al., 2020, , 70, 101–119
- Soszyński I., et al., 2022, , 72, 245–256
- Soszyński I., et al., 2023, , 73, 105–126
- Sota A., Maíz Apellániz J., Walborn N. R., Alfaro E. J., Barbá R. H., Morrell N. I., Gamen R. C., Arias J. I., 2011, , 193, 24
- Sota A., Maíz Apellániz J., Morrell N. I., Barbá R. H., Walborn N. R., Gamen R. C., Arias J. I., Alfaro E. J., 2014, , 211, 10
- Southworth J., 2012, in Arenou F., Hestroffer D., eds, *Orbital Couples: Pas de Deux in the Solar System and the Milky Way*. p. 51–58 (arXiv:1201.1388), doi:10.48550/arXiv.1201.1388
- Southworth J., Bowman D. M., 2022, , 513, 3191
- Stassun K. G., et al., 2019, , 158, 138
- Stellingwerf R. F., 1978, , 224, 953–960
- Storrie-Lombardi M. C., Lahav O., Sodre Jr. L., Storrie-Lombardi L. J., 1992, , 259, 8P
- Szklenár T., Bódi A., Tarczay-Nehéz D., Vida K., Marton G., Mező G., Forró A., Szabó R., 2020, , 897, L12

- Szklenár T., Bódi A., Tarczay-Nehéz D., Vida K., Mező G., Szabó R., 2022, , 938, 37
- Sánchez-Sáez P., et al., 2021, , 161, 141
- Tan M., Le Q. V., 2019, arXiv e-prints, p. arXiv:1905.11946
- Tayler R. J., 1994, *The Stars, Their Structure and Evolution*
- Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, *Astronomical Society of the Pacific Conference Series Vol. 347, Astronomical Data Analysis Software and Systems XIV*. p. 29
- Tkachenko A., Pavlovski K., Serebriakova N., Bowman D. M., IJspeert L., Gebruers S., Southworth J., 2024, , 683, A252
- Torres G., Andersen J., Giménez A., 2010, , 18, 67–126
- Udalski A., 2003, , 53, 291–305
- Udalski A., Szymanski M., Kaluzny J., Kubiak M., Mateo M., 1992, , 42, 253–284
- Udalski A., Szymański M. K., Szymański G., 2015, , 65, 1–38
- Udalski A., et al., 2018, , 68, 315–339
- Valenzuela L., Pichara K., 2018, , 474, 3259–3272
- Waelkens C., 1991, , 246, 453
- Walborn N. R., 1972, , 77, 312–318
- Walborn N. R., 1973, , 78, 1067–1083
- Wang Y., Li Y., Song Y., Rong X., 2020a, *Applied Sciences*, 10
- Wang C., Langer N., Schootemeijer A., Castro N., Adscheid S., Marchant P., Hastings B., 2020b, , 888, L12
- Watson C. L., Henden A. A., Price A., 2006, *Society for Astronomical Sciences Annual Symposium*, 25, 47
- Weidner C., Vink J. S., 2010, , 524, A98
- Weis K., Bomans D. J., 2020, *Galaxies*, 8, 20
- Wenger M., et al., 2000, , 143, 9–22
- Wolf C. J. E., Rayet G., 1867, *Academie des Sciences Paris Comptes Rendus*, 65, 292–296
- Woosley S. E., Weaver T. A., 1986, , 24, 205

- Wozniak P. R., 2000, , 50, 421–450
- Woźniak P. R., et al., 2004a, , 127, 2436–2449
- Woźniak P. R., Williams S. J., Vestrand W. T., Gupta V., 2004b, , 128, 2965–2976
- Wright E. L., et al., 2010, , 140, 1868–1881
- Yu J., Huber D., Bedding T. R., Stello D., Hon M., Murphy S. J., Khanna S., 2018, , 236, 42
- Zari E., Rix H. W., Frankel N., Xiang M., Poggio E., Drimmel R., Tkachenko A., 2021, , 650, A112
- Zhang K., Bloom J. S., 2021, , 505, 515–522
- Zhang Q., Yang L. T., Chen Z., Li P., 2018, *Information Fusion*, 42, 146–157
- Zhao G., Zhao Y.-H., Chu Y.-Q., Jing Y.-P., Deng L.-C., 2012, *Research in Astronomy and Astrophysics*, 12, 723–734
- Zhu Y., Hu Z., Zhang Q., Wang L., Wang J., 2006, in McLean I. S., Iye M., eds, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 6269, Ground-based and Airborne Instrumentation for Astronomy*. p. 62690M, doi:10.1117/12.671124
- Zinnecker H., Yorke H. W., 2007, , 45, 481–563
- de Jong R. S., et al., 2016, in Evans C. J., Simard L., Takami H., eds, *Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI*. SPIE, p. 99081O, doi:10.1117/12.2232832, <https://doi.org/10.1117/12.2232832>
- de Jong R. S., et al., 2019, *The Messenger*, 175, 3–11
- de Mink S. E., Pols O. R., Hilditch R. W., 2007, , 467, 1181–1196
- de Mink S. E., Sana H., Langer N., Izzard R. G., Schneider F. R. N., 2014, , 782, 7