DATA SHIFT MITIGATION IN CLASSIFIERS OF VARIABLE STARS

Thesis · May 2025

DOI: 10.13140/RG.2.2.14504.28165

1 author:

Francisco Pérez-Galarce
University of the Americas
32 PUBLICATIONS **214** CITATIONS

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
SCHOOL OF ENGINEERING

DATA SHIFT MITIGATION IN CLASSIFIERS OF VARIABLE STARS

FRANCISCO JAVIER PÉREZ GALARCE

Thesis submitted to the Office of Graduate Studies in partial
fulfillment of the requirements for the degree of
Doctor in Engineering Sciences

Advisor:
KARIM PICHARA

Santiago of Chile, (December, 2024)

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING

DATA SHIFT MITIGATION IN CLASSIFIERS OF VARIABLE STARS

FRANCISCO JAVIER PÉREZ GALARCE

Members of the Committee:

KARIM PICHARA

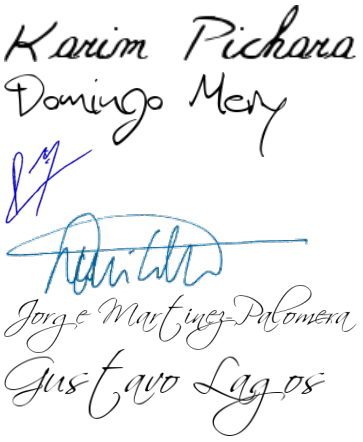
DOMINGO MERY

PABLO HUIJSE

MÁRCIO CATELAN

JORGE MARTINEZ-PALOMERA

GUSTAVO LAGOS



Thesis submitted to the Office of Graduate Studies in partial fulfillment of the requirements for the Degree Doctor in Engineering Sciences

Santiago de Chile, December, 2024

To Soledad and Francisca.

ACKNOWLEDGEMENTS

First and foremost, my heartfelt appreciation goes to my family, whose unconditional love and support have sustained me throughout this journey. Their belief in me, constant encouragement, and sacrifices have been the foundation of my resilience and success. I am eternally grateful to them for their unwavering faith in me.

Additionally, I extend my deepest gratitude to my advisor, Karim, whose guidance, patience, and expertise have been indispensable throughout this journey. Karim's insights and encouragement have shaped this research and inspired me to pursue excellence in all facets of my work. His tireless support and mentorship have been pivotal to my growth as a researcher and person. I am profoundly thankful for his dedication and the invaluable lessons learned under his tutelage.

I am also immensely grateful to my labmates for their camaraderie, support, and shared wisdom. Our lab meetings, brainstorming sessions, and moments of solidarity have enriched my research experience immeasurably. I am fortunate to have worked alongside such talented and dedicated individuals. Each of them has contributed to my project meaningfully, and I am deeply appreciative.

I thank my committee members, research collaborators, and professors who have contributed significantly to my academic and personal development. Their feedback, encouragement, and rigorous scrutiny have refined my research and broadened my perspectives.

Furthermore, I would like to acknowledge the support from the National Agency for Research and Development (ANID), through the Scholarship Program/Doctorado Nacional/2017-21171036.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Currently, machine learning plays a crucial role in the automatic classification of variable stars. Numerous classifiers have been proposed, achieving high performance in classification metrics. However, these classifiers are affected by biases in the training data, limiting their effectiveness on unseen data. These limitations damage the accuracy of predictions, leading to multiple issues, such as the selection of incorrect models and overestimated performance. This thesis introduces methods to evaluate and train classifiers under data shift conditions.

Firstly, an informative marginal likelihood is explored to select RR Lyrae star classifiers, mitigating biases through the incorporation of deterministic rules based on physical parameters during estimation. Then, we propose to improve the training of a multilayer perceptron for classifying the same type of stars. Knowledge is injected through characteristic intervals of physical parameters that inform a regularizer. A two-step backpropagation algorithm integrates this knowledge into the neural network, minimising classification error and controlling the balance between learning from data and expert knowledge. Finally, using the same dual learning scheme, a self-regulated learning approach is proposed to train a convolutional neural network. This training scheme takes advantage of recent advances in generative models to create synthetic light curves during training.

In summary, this thesis advances the integration of expert knowledge into machine learning models for astronomy, providing new evaluation and training methodologies for variable star classifiers. These advancements focus on improving generalisation and prediction capabilities. Additionally, they offer novel and pertinent learning frameworks for future research.

Keywords: variable stars; data shift; supervised learning; expert knowledge; classification.

RESUMEN

En la actualidad el aprendizaje de máquinas desempeña un papel crucial en la clasificación automática de estrellas variables. Numerosos clasificadores han sido propuestos, alcanzado un alto rendimiento en las diferentes métricas de clasificación. Sin embargo, estos clasificadores se ven afectados por sesgos en los datos de entrenamiento, lo que limita su efectividad en datos no vistos. Estas limitaciones dañan la capacidad de generalización, generando múltiples problemas como la selección de modelos incorrectos y un desempeño sobreestimado. En este contexto, esta tesis introduce métodos para evaluar y entrenar clasificadores bajo condiciones de desplazamiento de datos.

En primer lugar, se explora una probabilidad marginal informativa para seleccionar clasificadores de estrellas RR Lyrae, mitigando los sesgos a través de la incorporación de reglas deterministas basadas en parámetros físicos durante su estimación. Luego, se propone mejorar el entrenamiento de un perceptrón multicapa para la clasificación del mismo tipo de estrellas. Se injecta conocimiento a través de rangos de parámetros físicos que informan a un regularizador. Un algoritmo de retropropagación en dos pasos fusiona este conocimiento en la red neuronal, minimizando el error de clasificación y gestionando el equilibrio entre el aprendizaje a partir de datos y el conocimiento experto. Por último, utilizando el mismo esquema dual de aprendizaje se propone un aprendizaje auto regulado para una red neuronal convolucional, que toma ventaja de los recientes avances en modelos generativos para la creación de curvas de luz sintéticas durante el entrenamiento.

En conclusión, esta tesis avanza en la integración del conocimiento experto en modelos de aprendizaje automático, proporcionando nuevas metodologías de evaluación y entrenamiento para clasificadores de estrellas variables. Estos avances mejoran las capacidades de generalización y predicción, ofreciendo adicionalmente novedosos y pertinentes esquemas de aprendizaje para futuras investigaciones.

Palabras Claves: estrellas variables; desplazamiento de datos; aprendizaje supervisado; conocimiento experto; clasificación.

LIST OF ABBREVIATIONS

AE	Autoencoder
ACC	Accuracy
ACEP	Anomalous Cepheid
ANN	Artificial neural network
ASAS	All Sky Automated Survey
ASAS-SN	All-sky automated survey for supernovae
AUC	Area under the curve
BGMM	Bayesian Gaussian mixture model
BLR	Bayesian logistic regression
BLR-IP	Bayesian logistic regression with informative priors
BMS	Bayesian model selection
CEP	Classical Cepheid
CNN	Convolutional neural network
CoRoT	Convection, Rotation and planetary Transit survey
CV	Cross-validation
DL	Deep learning
DPV	Double periodic variable
DR	Deterministic rules
DSCT	Delta Scuti variables
DN	Dwarf nova
ECL	Eclipsing binary
FATS	Feature analysis for time series
F1	F1 score metric. Harmonic mean of precision and recall rates
GAN	Generative adversarial networks
GRU	Gated recurrent unit
HR	Hertzsprung-Russel
IWCV	Importance weighted cross validation
KNN	k nearest neighbours

LINEAR	LIncoln Near-Earth Asteroid Research survey
LPV	Long-period variable star
LR	Logistic regression
LSTM	Long short-term memory
MAP	Maximum a posteriori
MACHO	MAssive Compact Halo Object catalogue
MCMC	Markov chain monte Carlo
MC	Monte Carlo
ML	Machine learning
MLP	Multi layer perceptron
OGLE	Optical gravitational lensing experiment
PCA	Principal component analysis
PanSTARR	Panoramic Survey Telescope and Rapid Response System
PELS-VAE	Physical enhanced latent space variational autoencoder
RCB	R CrB variable star
ReLu	Rectified linear unit
RF	Random forest
RNN	Recurrent neural network
ROC	Receiver operating characteristic
ROC-OVR	Receiver operating characteristic in a one-vs-rest approach
ROC-OVO	Receiver operating characteristic in a one-vs-one
RRLYR	RR Lyrae variable star
RR_i	RR Lyrae i-type, $i \in \{ab, c, d, e\}$
SNR	Signal-to-noise ratio
SVM	Support vector machine
T2CEP	Type II Cepheids star
TC	True Class
tCNN	Temporal convolutional neural networks
VAE	Variational autoencoder

VVV	VISTA Variables in the Vía Láctea ESO public survey
WISE	Wide-field Infrared Survey Explore

1. CHAPTER I. INTRODUCTION

Machine learning models have been extensively applied to classify variable stars in recent decades. Each variable star exhibits a distinctive variability pattern, and identifying the type of a variable star based on its variability pattern is both an engaging and challenging problem. These variability patterns are analysed using light curves, which are graphs that depict an object's brightness over time. The challenge is due to the different sources and variety of patterns, which may be exogenous (e.g., planetary systems) or endogenous (e.g., eruptions). Additionally, certain characteristic patterns can overlap among the dozens of classes and subclasses of variable stars (Samus et al., 2017). Overcoming this challenge is crucial for astronomers due to the significance of variable stars. Variable stars are crucial celestial objects, mainly because some (i.e., RR Lyrae and Cepheid) are used to calculate reliable distance estimations (Beaton et al., 2016).

From a machine learning viewpoint, the main effort has been focused on developing classifiers (Debosscher et al., 2007, 2009; Pichara et al., 2012; Mackenzie et al., 2016; Benavente et al., 2017; Narayan et al., 2018; Carrasco-Davis et al., 2019; Aguirre et al., 2019; Becker et al., 2020) and new alternatives to represent the celestial objects, i.e., human-based features and automatic features as well, (Nun et al., 2015; Kim and Bailer-Jones, 2016; Valenzuela and Pichara, 2017; Naul et al., 2018; Donoso-Oliva et al., 2023). The amount of data generated by modern telescopes has encouraged using complex models to automate certain classification tasks. Examples of relevant machine learning models proposed for variable stars classification include support vector machine (SVM; Debosscher et al., 2009; Benavente et al., 2017), random forest (RF; Richards et al., 2011; Bloom et al., 2012; Pichara et al., 2012; Benavente et al., 2017; Narayan et al., 2018), convolutional neural networks (CNN; Aguirre et al., 2019), recurrent neural networks (RNN; Becker et al., 2020), autoencoders (AE; Naul et al., 2018), and even ad-hoc neural networks for periodic time-series have been proposed (Zhang and Bloom, 2021).

Despite these great advances, we can visualise at least two drawbacks in the training data; first of all, the community has not used fully representative surveys to train these models, and some classes in the training data have very few labelled objects. Both situations hinder the model assessment procedure; hence, the performance of these models is not clear when tested in objects out of the training data. Even more important, we could use wrong models to label new training data, generating a cascade effect. It has given rise to the next questions, *Can we improve the model assessment process? Can we inject expert knowledge to mitigate these biases? Can we train more reliable classifiers?*

Several papers have commented about the biases in training data of variables stars (Debosscher et al., 2009; Richards, 2012; Masci et al., 2014); however, a solution for better training and assessing classifiers in this scenario has not been proposed. Bias in our data means that there is a difference between the joint distribution of our labelled data $\mathcal{D}^S$ and the joint distribution of the population $\mathcal{D}^P$. The problem arises because current classifiers are trained with a subset of $\mathcal{D}^S$, and after that, the performance is evaluated using the complement from the same $\mathcal{D}^S$ (the testing set) typically in a cross-validated (CV) scheme.

Those biases stem from several sources; they can be linked with human tasks and the technical characteristics of the telescopes. The first one is associated with the labelling process because astronomers label a type of star more frequently when it is easier to define; a good discussion about this systematic bias in the astronomical labelling process was presented in Cabrera et al. (2014). The mechanical design of receptors generates another type of bias, specifically by the range where the signal can be processed; e.g., when the distance increases, less luminous objects are more challenging to see (Richards, 2012). Moreover, the fast development of technologies accelerates the obsolescence of the model; hence, we can not apply trained models to new surveys, or we do not have enough confidence about the error metrics in these newer catalogues; the former problem is also known as domain adaptation (or transfer learning), and it was in-depth discussed for variability surveys by Benavente et al. (2017). Lastly, observation scheduling of telescopes can also

induce data shift. Selection biases emerge when specific sky regions are preferentially observed, potentially leading to covariate shifts if these regions differ from others in critical characteristics. Variations in observation times, cadence, and environmental conditions can also alter the data collected, affecting the distribution and types of celestial objects observed.

To analyse the effect of these biases, we divide them into two categories: biases in features and biases in the class representation. The bias in features represents a difference in the joint feature distribution between $\mathcal{D}^S$ and $\mathcal{D}^P$ without considering the class representation; zones of the space of features can express this without labelled objects or over-representation of other zones what typically generates a change in the relevance of these zones in the assessment process. The bias in the class proportion is related to the difference of the class representation between $\mathcal{D}^S$ and $\mathcal{D}^P$; it comes from training sets that contain classes with very few labelled objects. However, these classes can have a higher representation in real scenarios. It generates additional challenges due to the difficulty of validating and learning with available data.

This problem, also known as data shift, which is latent in many real-world problems, has been studied extensively by the machine learning community; Quiñonero-Candela et al. (2009) provide an excellent introduction to the subject with a description of the possible types of data shift, namely, target shift, covariate shift and conditional shift. The covariate shift considers $p^S(\mathbf{x}) \neq p^P(\mathbf{x})$, and $p^P(y|\mathbf{x}) = p^S(y|\mathbf{x})$, where $p^S(\mathbf{x})$ and $p^P(\mathbf{x})$ are the density probability distributions of features in training and testing sets, respectively; and, $p^P(y|\mathbf{x}) = p^S(y|\mathbf{x})$ are conditional distributions of the label given the features, in training and testing sets, respectively. In other words, we are in the presence of a covariate shift problem when features in the testing set, e.g. period or amplitude, have a different joint density distribution with respect to the training set, even considering a representative conditional distribution. The target shift assumes $p^P(y) \neq p^S(y)$, but $p^S(\mathbf{x}|y) = p^P(\mathbf{x}|y)$. It is to say that a target shift involves a mismatch between the label proportion of each class in the training set and the testing set. The conditional shift is characterised by $p^S(y|\mathbf{x}) \neq$

$p^P(y|\mathbf{x})$ and $p^P(y) = p^S(y)$; this type of data shift is observed when the relationship of one or all the classes change with respect to the features; for example, for the same feature values a star type can have a different probability value in the testing set from the training set. In the majority of scenarios, it is very complex to divide these biases, and all of them, in conjunction, directly affect model performance, i.e., $p^S(\mathbf{x}, y) \neq p^P(\mathbf{x}, y)$, it is to say, there is a difference in the joint probability distribution of labels and features between training and testing sets.

Even with this latent problem, little effort has been made to study metrics and validation strategies to evaluate the performance of light curve classifiers. Besides, providing more accurate metrics for assessing models in a scenario where we can only partially trust the data seems challenging. A natural frame to face the problems as mentioned earlier is Bayesian modelling, which is being increasingly used in different astronomy fields, such as the comparison of astrophysical models (Ford and Gregory, 2006) or making predictions of properties of celestial objects (Das and Sanders, 2018). In order to improve the model assessment task, this thesis, as the first step, proposes a novel pipeline to evaluate variable star classifiers in training sets that contain several problems like bias and a few labelled objects. The methodology is based on Bayesian machine learning, which allows us to incorporate astronomical knowledge in the model assessment process. Specifically, this approach is based on the Bayesian model selection scheme (Murray and Ghahramani, 2005), which embodies desirable properties such as Bayesian Occam's razor, consistency, and comparability.

Moreover, few papers have proposed scalable alternatives to mitigate the impact of the biases mentioned above in the model training phase. The need for solutions to this problem stems from various sources, such as the difficulty of controlling and injecting knowledge into state-of-the-art machine learning models such as artificial neural networks (ANNs). Blending human knowledge into artificial neural networks and broader machine learning systems is a complex task that can lead to better, more reliable classifiers when dealing with the data shift problem. This area has been explored by many experts, like

Von Rueden et al. (2021), Deng et al. (2020), and Borghesi et al. (2020), which highlighted the complexity of implementing this intuition. One alternative is creating synthetic data and injecting these objects during the training, which is seen as a way to improve data quality and fix those biases. However, understanding how the training set is biased remains challenging, and consequently, how to inject suitable synthetic objects without introducing new biases. As a second and third step, this thesis offers two alternatives for injecting expert knowledge when artificial neural networks are trained; the first approach proposes injecting deterministic rules into a multi-layer perceptron training, and the second method seeks to improve the convolution neural network reliability through synthetic light curves produced by a conditional variational autoencoder.

1.1. Hypothesis and objectives

This thesis is focused on providing methods that improve the reliability of periodic variable star classifiers when the data shift problem generates biased trained models. Three main hypotheses are proposed, covering the phases of model selection and model training as follows:

1.1.1. Hypotheses

- H1: Injecting simple human knowledge into the marginal likelihood estimation can reduce biases within a Bayesian model selection framework, thereby providing a robust method for evaluating variable star classifiers in the presence of data shift issues.
- H2: A regularisation approach informed by expert knowledge and an ad-hoc training strategy can mitigate the data shift problem, resulting in more reliable classifiers.
- H3: Incorporating samples on demand from a generative model of variable stars during the training phase can help address under-represented areas in the physical parameters, enhancing the model performance and reducing the impact of shifted training data.

1.1.2. Objectives

1.1.2.1. General objective

Develop and implement methods that enhance the reliability of periodic variable star classifiers, particularly addressing the challenges posed by the data shift problem, using expert knowledge from physical parameters.

1.1.2.2. Specific objectives

O1: Implement a method that integrates simple human knowledge into the marginal likelihood estimation within a Bayesian model selection framework, aiming to reduce biases and provide a robust evaluation mechanism for features-based binary classifiers of RR Lyrae stars under data shift conditions.

O2: Develop a regularisation approach, including expert knowledge, combined with a specific training strategy, to mitigate the effects of the data shift problem and achieve more reliable binary classifier of RR Lyrae stars.

O3: Incorporate generative modelling for periodic variable stars in the training process to produce synthetic samples, aiming better to represent under-represented areas in the physical parameters space, thereby improving the multi-class classifier reliability and robustness against data shifts.

1.2. Contributions

In this section, the thesis details significant advancements in variable star classification, achieved by integrating domain-specific knowledge to counteract data shift problems.

1.2.1. Informative Bayesian model selection for RR Lyrae star classifiers

One natural framework with which to address the data shift problem is Bayesian modelling, which has been increasingly used in different fields of astronomy, such as to compare astrophysical models (Ford and Gregory, 2006) or make predictions on the properties of celestial objects (Das and Sanders, 2018). A novel pipeline for evaluating Bayesian logistic regressions in biased training data is proposed to improve the model assessment task. The methodology, based on Bayesian machine learning, allows the incorporation of astronomical knowledge into the model assessment process. In particular, this approach exploits the robust Bayesian model selection (BMS) scheme (Murray and Ghahramani, 2005), which embodies desirable properties such as Bayesian Occam's razor, consistency, and comparability (Myung and Pitt, 1997).

The Bayesian model selection framework present a great computational challenge for most statisticians and data scientists. However, even the powerful marginal likelihood cannot assess models correctly when the training data are biased. This strategy exploits expert knowledge to address these biases by incorporating informative priors in the marginal likelihood estimation of RR Lyrae stars classifiers. The proposed methodology is divided into three stages; first, a method to represent the prior knowledge using deterministic rules (DRs) based on physical-based features, such as period and amplitude, is designed. In the second stage, samples from the posterior distribution using these informative priors are generated. Samples from the posterior distribution is suitable because it ensures that samples are drawn from regions with high values in both the likelihood function and the prior distribution. Moreover, astronomical knowledge through the effect of the priors in the posterior distribution can be incorporated. Finally, in the third phase, the marginal

likelihood is estimated using the approximated bridge sampling estimator (Overstall and Forster, 2010; Gronau et al., 2017).

1.2.2. Informative regularisation for a multi-layer perceptron RR Lyrae classifier under data shift

Based on the previously presented contribution, which demonstrates that simple deterministic rules can improve the model selection phase, it is proposed to use this straightforward representation of expert knowledge to design a novel regularisation strategy for ANNs.

This method proposes a scalable and easily adaptable regularisation scheme to incorporate astronomical knowledge into ANNs. This approach provides an alternative for training more reliable RR Lyrae classifiers, even in the presence of data shift issues. An ad-hoc modelling and training procedure is designed to enforce effective informative regularisation. This algorithm employs a double error propagation in each epoch, using two disjoint sets of weights (masks). The first propagation aims to reduce the classification error, while the second spreads the informative regularisation throughout the neural network. The training procedure includes an initial phase where weights are assigned to each loss function.

1.2.3. A self-regulated convolutional neural network for classifying variable stars

The third contribution proposes a novel approach to train more reliable variable star classifiers by leveraging recent advancements in synthetic data generation based on deep learning models. We propose integrating a generative model with a classifier in a cooperative framework; our approach dynamically enhances the learning process with synthetic examples in under-represented areas. During the classifier training, the synthetic samples, which are focused on mitigating biases and imbalance problems, are initially obtained from the stellar physical parameter space, including effective temperature, period, metallicity, absolute magnitude, surface gravity, and radius. These samples are then processed

by a trained physics-enhanced latent space variational autoencoder (PELS-VAE), which returns a synthetic light curve. We highlight that sampling from the physical parameter space, which is a low-dimensional space, allows us to manage the over/under-represented zones when generating new light curves. Our method adjusts the classifier training trajectory through the injection of new objects from the generative model. We propose five policies for defining the number of samples for each class, some aimed at populating classes where the confusion is most significant according to the confusion matrix in the current training epoch. A mask-based training scheme is included to avoid competition between real and synthetic data. We also provide a set of experiments, considering two types of data shift, that assess the classifier performance under variations in the signal-to-noise ratio and the sequence length, highlighting where synthetic samples are most relevant for reducing data shift and the impact of class imbalance. Finally, we demonstrate that our synthetic light curves can assist in training more reliable classifiers and optimising hyperparameters, which remains a challenging task in current DL-based architectures.

1.3. Variable stars

Variable stars are stellar objects whose brightness changes over time. This variability is attributed to diverse factors, both intrinsic and extrinsic, to the stars themselves. Periodic variable stars display changes in brightness that follow a regular, repeating pattern. This record of information can be represented like a time series; in the astronomical context, the term *light curve* is preferred. These fluctuations in luminosity are often the result of internal physical processes within the stars, e.g. pulsating variable stars (Catelan and Smith, 2015) or extrinsic phenomenon, e.g. cataclysmic variability (Smith, 2006). However, extrinsic factors, such as orbital dynamics, can also lead to periodic differences in brightness. For instance, consider a binary star system where two stars orbit each other.

Figure 1.1 presents the light curve of the star OGLE-BLG-RRLYR-08634, illustrating the variation in magnitude over time. Subsequently, Figure 1.2 shows the folded light curve of the same star, where the observational data are folded into phases based on the

star's period. This folded approach is instrumental in discerning the repetitive and characteristic pattern of brightness variation, enabling a more detailed analysis of the star's variability properties.

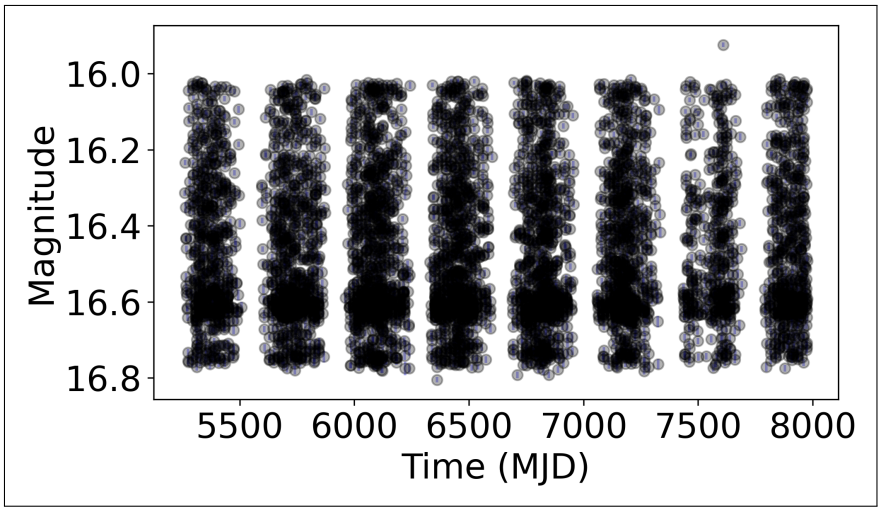


Figure 1.1. RR Lyrae star ID OGLE-LMC-RRLYR-08634.

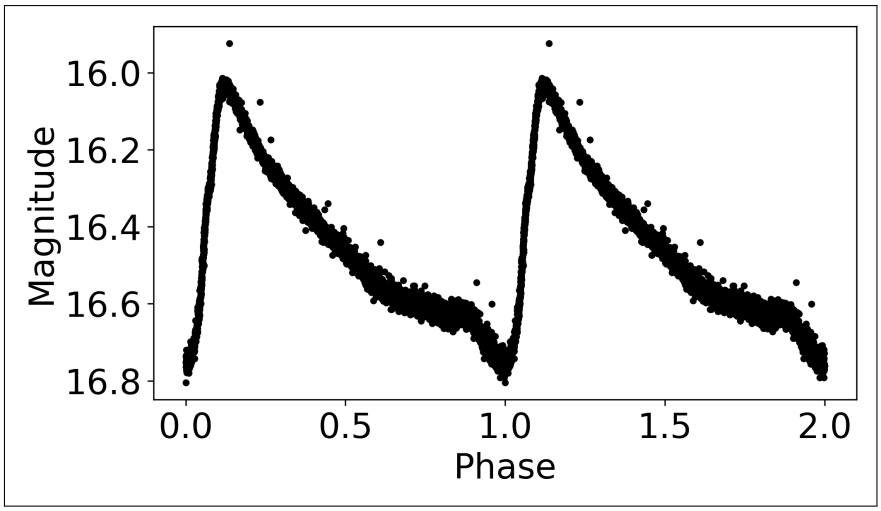


Figure 1.2. Folded light curve for RR Lyrae star ID OGLE-LMC-RRLYR-08634.

The study of variable stars contains various classifications, with dozens of types and sub-types, herein also referred to as classes and subclasses, identified in the field (Samus et al., 2017). This reflects the wide variety of mechanisms and phenomena that can induce stellar variability. Researchers organise these stars using a taxonomy based on their

characteristics and behaviour. A well-known taxonomy can be seen in Figure 1.3, which presents an overview of the different types of variable stars and other variable phenomena. First, this variability tree separates variable stars into extrinsic and intrinsic types based on the cause of their variability. Extrinsic variables change brightness due to external factors, such as eclipses in binary systems (Kallrath et al., 2009). On the other hand, intrinsic variables vary due to internal processes, for example, eruptive stars with violent outbursts or explosions (Griffiths, 2018), cataclysmic variables involving accretion-induced outbursts (Smith, 2006), pulsating stars that rhythmically expand and contract (Catelan and Smith, 2015), and secular variables experiencing long-term changes due to evolutionary processes. This classification framework is crucial for astronomers to understand and organise the diverse behaviours observed in stellar variations.

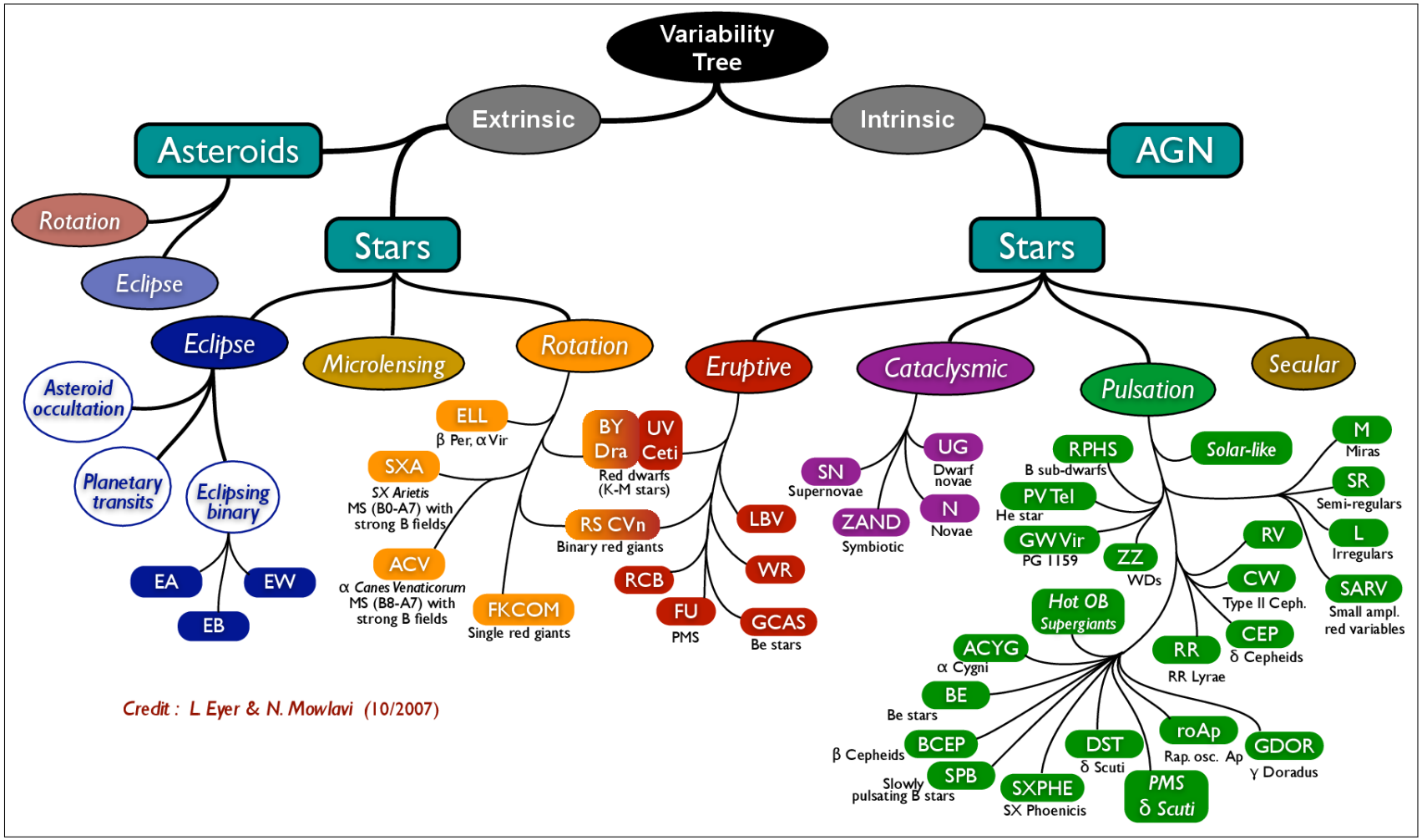


Figure 1.3. The variability tree proposed by Eyer and Mowlavi (2008) and updated by Gaia Collaboration et al. (2019)

This thesis examines five distinct classes of variable stars: RR Lyrae, Cepheid, long period variables, eclipsing binaries, and delta Scuti. The initial focus is developing binary classifiers tailored for RR Lyrae stars in the first and second contributions. Subsequently, the third contribution of the thesis involves the integration of these five periodic classes into a comprehensive multi-class classifier.

RR Lyrae stars: Variable stars well-known for their advanced age and low metallicity, [Fe/H], are found in the dense confines of globular clusters. Characterised by shorter periods, commonly staying less than a day, they are essential for distance estimation within the Milky Way and its surrounding satellite galaxies. There are well-identified sub-types within the RR Lyrae class, such as type-ab RR Lyrae stars, RRab, have periods ranging from 0.2 to 1 days; type-c RR Lyrae stars, RRc, have periods ranging from 0.2 to 0.5 days; and type-d RR Lyrae stars, RRd, have periods ranging from 0.2 to 0.7 days. RR Lyrae stars exhibit distinct characteristics in their periodic light variations and fundamental physical properties. The light curves of RR Lyrae stars, marked by their specific shape and periodicity, indicate their variability type. Moreover, effective temperatures range from approximately 6,000 to 7,250 Kelvin, and diameter is typically 4-6 $R_{\odot}$ (Catelan and Smith, 2015). The surface gravity values typically fall within a narrow range, suggesting a specific strength of the gravitational force at their surfaces, which affects the star's atmospheric pressure. Lastly, RR Lyrae stars have a predictable absolute magnitude from 40 to 50 times brighter than the Sun, 0.19 to 3.7 in band G (Eyer et al., 2023), making them recognisable as standard candles for distance measurement. The Hertzsprung-Russell (HR) Diagram presented in Figure 1.4 is a fundamental tool in understanding stellar evolution and positions stars according to their luminosity and temperature. On the HR diagram, RR Lyrae stars are found on the horizontal branch, a stage that indicates helium core burning. Figure 1.2 provides a common variability pattern for RR Lyrae stars, see more example in Figure 1.5.

Cepheid stars: Luminous super-giant stars are characterised by periodic changes in brightness, with variability that is predictable by the period-luminosity relation, which is caused

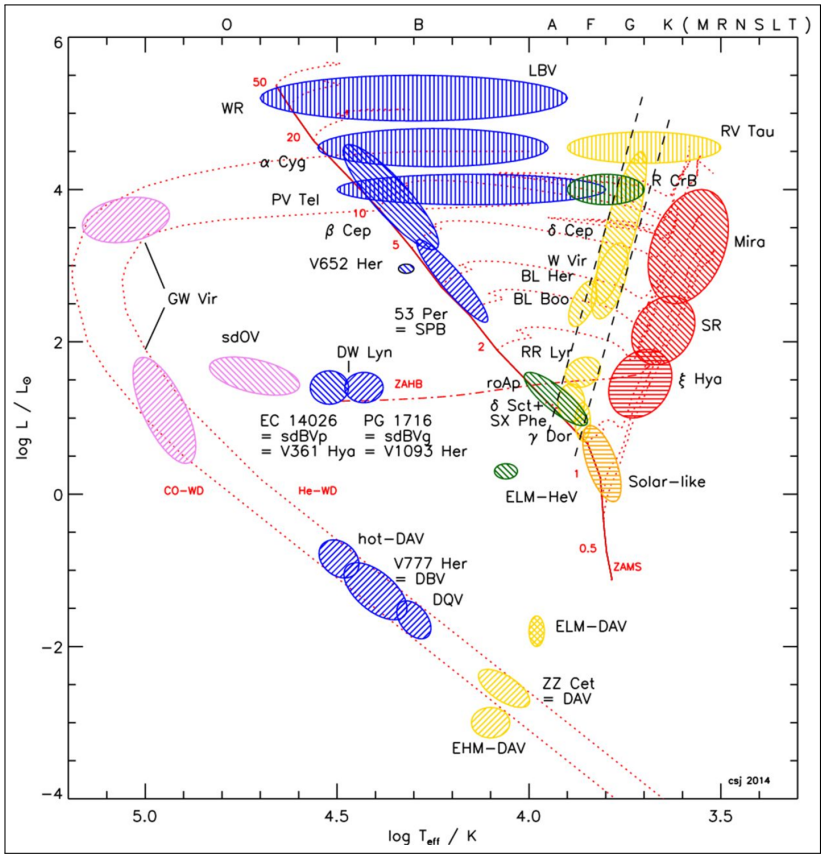


Figure 1.4. Luminosity-effective temperature (or H–R) diagram from Jeffery and Saio (2016).

by internal physical mechanisms. This significant correlation entitles Cepheids as standard candles for gauging cosmic distances. Cepheids are categorised into two main types: classical Cepheids, with periods ranging from 1 to 100 days, and Type II Cepheids, with periods from 20 to 200 days (Catelan and Smith, 2015). Cepheids have effective temperatures from 5,000 to 7,000 Kelvin (Jayasinghe et al., 2021; Groenewegen, 2020; Kovtyukh et al., 2023) and radii extending from 20 to 400 times that of the Sun (Gieren et al., 1998). Surface gravity, $\log(g)$, in these stars is typically between 0.7 to 2.5, (Jayasinghe et al., 2021), and their absolute magnitudes in band G span from -2.8 to 1.1 (Eyer et al., 2023). These stars' predictable brightness variations support their role as crucial tools in astronomy, allowing the measurement of extensive cosmic distances. Figure 1.5 illustrates examples for variability patterns of Cepheid stars.

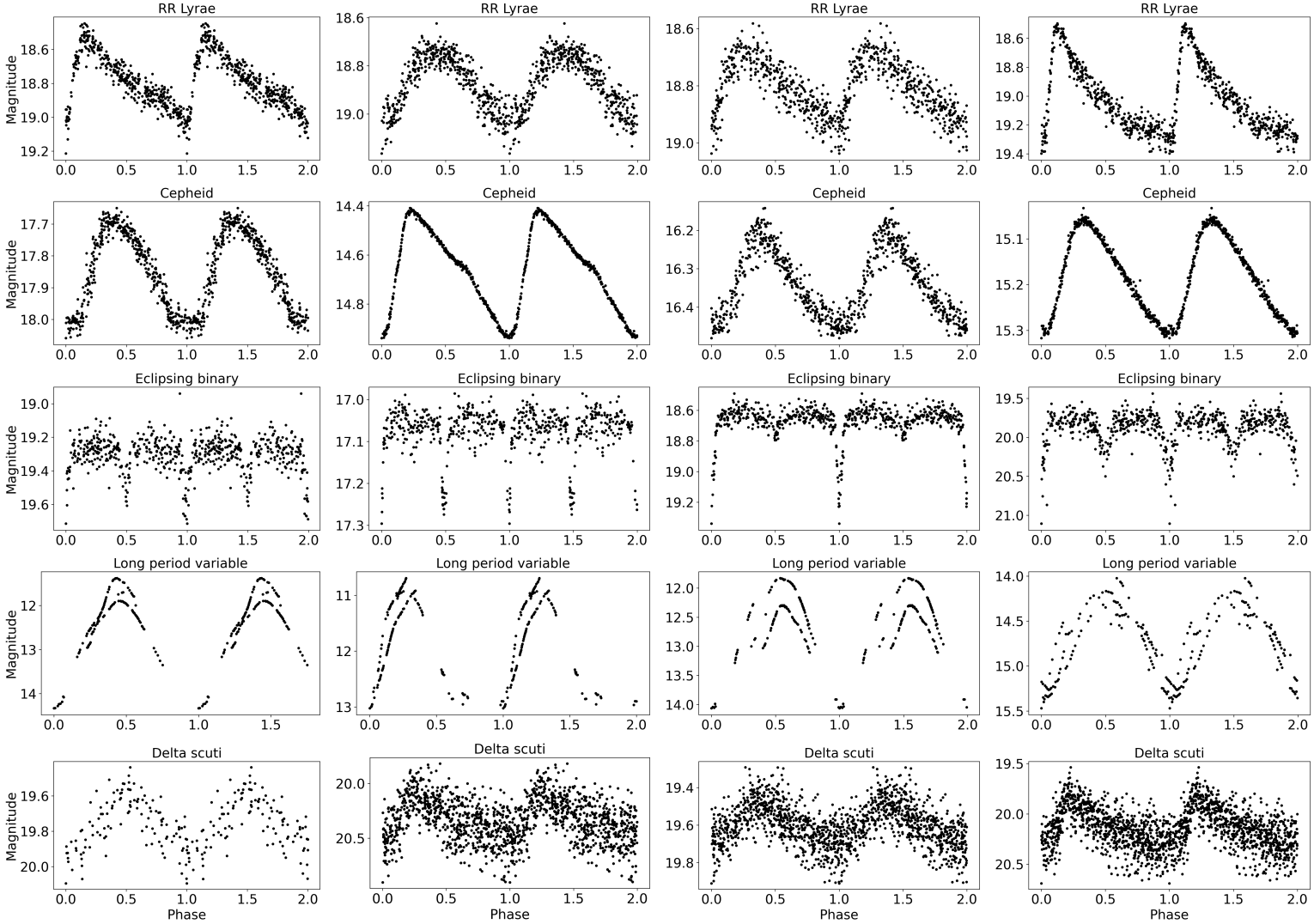


Figure 1.5. Sample of folded light curves for star types studied in this thesis.

Eclipsing Binaries: Eclipsing binaries are interesting stellar systems where two stars orbit each other closely, with their orbital plane oriented so that they periodically cover each other when viewed from Earth, causing variations in observed brightness. These systems are vital for deriving stellar properties like mass, radius, and temperature acquired through analysing their light curves during eclipse events. The sub-types are highly related to the Roche lobe in the stars of these systems, which is the region around a star in a binary system within which orbiting material is gravitationally bound to that star. This star type includes detached systems, where neither component fills its Roche lobe (Ivanova, 2014); semi-detached systems, where one component fills its Roche lobe while the other does not; and over-contact systems, where both components exceed their Roche lobes, each reflecting different orbital and physical interactions (Kallrath et al., 2009). The periods of these systems vary significantly, from a few hours to several days or even longer (Copperwheat et al., 2011). Eclipsing binaries have effective temperatures typically between 3,000 and 10,000 Kelvin (Armstrong et al., 2013). Their sizes can vary significantly, with radii from a fraction of up to several times that of the Sun. Moreover, the luminosity of eclipsing binaries can also vary widely, highlighting their importance in the study of stellar development and binary star dynamics. The analysis of their light curves provides essential data for understanding the intricate nature of these systems. Eclipsing binaries can appear in various parts of the HR Diagram, depending on the features of the separate stars in the binary system. Each star in a binary pair has properties, such as luminosity and temperature, placed in a specific location in the HR Diagram. Figure 1.5 displays examples of variability patterns for eclipsing binaries.

Long period variables (LPVs): These stars are known for their irregular and significant brightness variations, typically extending over periods ranging from several weeks to years. Some relevant types of LPVs are Mira, semi-regular, OGLE small amplitude red giants (OSARGs), and long secondary period stars. Mira variables are characterized by pulsation periods that range from about 100 to 1,000 days, exhibiting regularity in their brightness variations. Semi-regular variables show less predictable variations, with their periods ranging from tens to hundreds of days (Trabucchi et al., 2021). OSARG stars

have periods ranging from a few days to around 100 days. Long Secondary Period variables exhibit periods ranging from 200 to more than 1,000 days. As observed in Figure 1.4, these stars exhibit a cool nature with effective temperatures between 2,000 and 4,000 Kelvin (Feast, 1996). Their sizes are immense, with radii extending from tens to hundreds of times larger than the Sun, reflecting their advanced evolutionary stages (Feast, 1996). The surface gravity of these stars varies from approximately 0.5 to 2 times that of the Sun. The HR Diagram typically places LPVs on the asymptotic giant branch, highlighting their status as evolved stars often experiencing substantial mass loss. Figure 1.5 provides examples of typical variability patterns for LPVs.

Delta Scuti: delta Scuti stars are known for their short-period pulsations, both radial and non-radial, typically staying from minutes to a few hours (Handler, 2009). These stars are prevalent in various astronomical environments, such as young open clusters and the general field population of the Milky Way. Delta Scuti stars exhibit a range of effective temperatures from 6,000 to 9,000 Kelvin, suggesting a relatively hot nature (Uytterhoeven et al., 2011). In terms of size, these stars have radii that vary from a few to several times that of the Sun and their $\log g$ varies from two to five (Uytterhoeven et al., 2011). The luminosity of delta Scuti stars vary considerably, highlighting their heterogeneity and the complexity of their stellar processes. Delta Scuti stars are instrumental in asteroseismology, as their pulsations provide insights into stars' internal structures and evolutionary states. They inhabit a middle position in the HR diagram, usually found on or near the main sequence, bridging the gap between low-mass and high-mass stars (see Figure 1.4). This location in the HR diagram makes them helpful in investigating the transitionary phases of stellar evolution, as this class of pulsators contains stars of different ages and at different evolutionary stages. Figure 1.5 illustrates examples for the typical variability pattern of delta Scuti stars.

1.4. Related work

This section is divided into five subsections. Each subsection covers state-of-the-art papers about variable star classification using different lenses. Section 1.4.1 studies the state-of-the-art variable star classifiers, emphasising research addressing underlying biases and how these approaches select and compare models.

Then, Sections 1.4.2 and 1.4.3 provide a narrative on how automatic classifiers have evolved during the last two decades, considering that the machine learning contributions can be divided into the following two phases. The first phase was mainly based on machine learning models supported by feature engineering, and the second phase was driven mainly by deep learning models. Consequently, in Section 1.4.2, related work on variable stars classification with shallow models is summarised. Then, in Section 1.4.3, deep learning (DL) approaches are described.

To continue, given that RR Lyrae stars are an essential study case in the proposed methods, a more focused discussion of RR Lyrae classifiers is provided in Section 1.4.4.

After that, due to synthetic light curves providing a path to incorporate new information during the training step and consequently mitigate the biases in the training data, Section 1.4.5 presents a review of the state-of-the-art papers on generative modelling of variable stars setting the base knowledge to include them into the methods here proposed.

Finally, Section 1.4.6 briefly discusses how Bayesian data analysis has been used in astronomy. This section summarises how Bayesian approaches have been employed to select models, provide more reliable predictions and incorporate human knowledge in the modelling. This section provides the context for the first proposal contained in this thesis.

1.4.1. Classification of variable stars under bias

Several papers on the automatic classification of variable stars have aspired to address the data shift problem from different perspectives. However, none has focused on model

selection strategies. Over a decade ago, Richards et al. (2011) proposed several strategies to improve the training data. For example, they designed an active learning method and presented an importance-weighted CV method to avoid under-represented zones of feature space. However, metrics to compare models in these contexts were not analysed in-depth. Masci et al. (2014) proposed a random forest (RF) classifier, which was trained with a labelled set from the Wide-field Infrared Survey Explorer (WISE; Wright et al., 2010), within an active learning approach. This classifier was able to improve the training data and mitigate the biases. This RF approach outperformed SVM, k-nearest neighbors (KNN), and ANNs by using a CV method to estimate accuracy (ACC), which is defined as the proportion of correctly classified instances among the total instances.

Vilalta et al. (2013) addressed the issue of dataset shift in classifying Cepheid variable stars from galaxies at different distances. They introduced a novel two-step approach to align the distributions of period and apparent magnitude features between source and target datasets, enhancing coherence without modifying the trained model or re-weighting samples. Specifically, they introduced a parameter δ, which adjusts the apparent magnitude in the test dataset to match the distribution of the training dataset. This adjustment is estimated using maximum log-likelihood with respect to the probability density of the training data. Once the datasets are aligned, traditional classifiers are employed, including ANNs, SVM, RFs, and decision trees. This method results in classification ACC comparable to scenarios where labels are available for the target data.

Benavente et al. (2017) proposed a full probabilistic model to address the domain adaptation problem. This model could transfer knowledge (feature vectors) among different catalogues. It was able to manage the covariate shift and improve the cross-validated F_1 score. F_1 score is a popular metric, defined as the harmonic mean of two more basic metrics: precision and recall, where precision is the proportion of true positives among all predicted positives, and recall is the proportion of true positives among all actual positives. A Gaussian mixture model representing each catalogue (source and target) and a mixture of linear transformations (translation, scaling and rotation) were applied.

Aguirre et al. (2019) designed a convolutional neural network that was able to learn from multiple catalogues, outperforming a features-based RF. To manage the imbalanced classes, Aguirre et al. (2019) proposed a novel data augmentation scheme which creates new light curves by sampling observations from real objects. They used two parameters, *burning* and *step*. The *burning* parameter indicates the number of observations to be discarded at the beginning of the light curve, while the *step* parameter defines the number of observations to be skipped. Using this procedure, they ensured variation in the sampled light curves.

Sooknunan et al. (2021) reported the relevance of a non-representative $\mathcal{D}^S$ when applying trained models in new telescopes. Moreover, they studied how the ACC metric decreases (training vs. real) when $\mathcal{D}^S$ is small. They used a few real objects to create the training data and generated synthetic light curves using a Gaussian process. Experiments with the next five classes of transients were conducted: active galactic nuclei, supernovae, X-ray binaries, γ-ray bursts and novae. The results showed that performance on new surveys improves when contextual information (such as object location) and multi-wavelength data are incorporated. To encourage the use of multi-wavelength information, they presented results with the optical telescope MeerLICHT (Bloemen et al., 2016) and the radio telescope MeerKAT (Booth and Jonas, 2012).

Naul et al. (2018) proposed using a recurrent AE to learn a variable star embedding from folded light curves. The measurement error in observations is used to weigh the reconstruction metric in the loss function for facing the underlying uncertainty in each light curve observation. Therefore, those observations with large measurement errors were less important. Subsequently, this embedding is used to classify through an RF classifier. The new representation is compared with two baseline sets of features (Richards et al., 2011; Kim and Bailer-Jones, 2016), being competitive with traditional approaches. It outperformed or was similar to the baselines in the LIncoln Near-Earth Asteroid Research survey (LINEAR; Sesar et al., 2013) and the MAssive Compact Halo Object catalogue (MACHO; Alcock et al., 1997).

Becker et al. (2020) presented an end-to-end scalable RNN light curve classifier capable of learning representations without human intervention or feature estimation, such as period estimation. Since the model does not require a period, it can work with both periodic and non-periodic stars. The researchers achieved competitive ACC in a shorter runtime than an RF based on handcrafted features. Furthermore, they discussed the comparison between biases in handcrafted features and those in deep-learning features, supporting the idea that deep-learning models can learn less biased features when applied to specific surveys.

Burhanudin et al. (2021) presented an RNN focused on handling the imbalanced problem observed in the Gravitational-wave Optical Transient Observer survey (Steeghs et al., 2022). The proposed RNN considers two types of inputs; on one hand, they include light curve information, i.e. time, magnitude and photometric error; on the other hand, they use contextual information, namely, the galactic coordinates expressed in degrees and the distance measured in arcseconds to the closest galaxy listed in the Galaxy List for the Advanced Detector Era catalogue (Dálya et al., 2018). The dataset set contains variable stars, active galactic nuclei and supernovae, having 99% of objects belonging to the majority class (variable stars). They highlight the crucial role played by loss functions, where a focal loss is proposed (Lin et al., 2017), and the contextual information, suggesting that the strength of a classification system lies not just in the structure of the model but equally in the approach to data handling.

To summarise, classifiers have a latent drawback related to the data shift problem in training sets. This issue has been faced a few times in variable star classification, but no definite solution (scalable and reliable for online classifiers) has been established. New methodological proposals should tackle both problems jointly to provide transferable classifiers. From the reviewed research articles, we can highlight that treating the data shift problem by improving data or creating synthetic data is a promising path to training more

reliable classifiers. We also conclude that most studies analysed herein have applied metrics based on the confusion matrix and have primarily utilised k-fold to validate the model performance and select models.

1.4.2. Shallow classifiers of variable stars

During the last two decades, the use of machine learning models to study and classify variable stars has increased significantly. The machine learning contributions can be divided into the following two phases. In the first phase, several research articles exploited and applied traditional classifiers, such as tree-based models, as well as logistic regressions (LRs), shallow ANNs, support vector machine, Gaussian mixture model classifiers, Bayesian networks and KNN (Debosscher et al., 2007, 2009; Richards et al., 2011; Pichara et al., 2012; Nun et al., 2014, 2015; Mackenzie et al., 2016; Benavente et al., 2017). The feature engineering that supported these applications showed how they can be used to speed up, for example, the detection of peculiar objects (Nun et al., 2014); the automatic classification of catalogues (Pichara et al., 2012; Nun et al., 2014; Pichara et al., 2016); and the modelling of domain adaptation on different telescopes (Benavente et al., 2017). Some specialised computational packages were designed to extract features from light curves (Hartman, 2012; Nun et al., 2015; VanderPlas, 2016; Naul et al., 2016; Barbary et al., 2016).

Due to the primary objective of these approaches being the extraction of knowledge from labelled objects, most of them only focused on the model performance within catalogues; consequently, they did not worry about the data shift problem of these labelled objects. Some studies contributing to the data shift problem during this phase were Richards et al. (2011) and Masci et al. (2014). Richards et al. (2011) highlighted the problem and proposed specific strategies (active learning and importance-weighted cross-validation) to improve the training data. Masci et al. (2014) also proposed an active learning method to train an RF with a labelled set from WISE (Wright et al., 2010). Both approaches are

based on active learning, a human-in-the-loop approach, where an expert provides prioritised object labels for improving the classification performance (Settles, 2009). This method can be highly time-consuming and prone to error, as these experts may also add biases to the labelled data. Hence, a complementary approach to mitigate those biases is needed.

1.4.3. Classification of variable stars with deep learning

Later, the availability of more extensive catalogues of light curves, in addition to certain breakthroughs in machine learning and artificial intelligence (e.g., deep learning), led to the incorporation of more complex machine learning models for variable stars classification (Narayan et al., 2018; Naul et al., 2018; Carrasco-Davis et al., 2019; Aguirre et al., 2019; Becker et al., 2020; Jamal and Bloom, 2020; Zhang and Bloom, 2021). Despite this remarkable progress, there is not contributions for mitigating the data shift problem in deep ANNs.

To the best of our knowledge, the first DL model proposed to classify light curves of variable stars was proposed by Mahabal et al. (2017), which presented an application of CNNs for classifying light curves from the Catalina Real-Time Transient Survey (CRTS; Drake et al. (2009)). In their method, light curves are converted into two-dimensional images, termed $\Delta m \Delta t$-images, characterising the change in magnitude (Δm) across varying time intervals (Δt); these $\Delta m \Delta t$-images are inputs for CNNs to classify. They achieved an 83% ACC rate, estimated by a five-random train-test split approach, in classifying periodic variable stars into seven classes, obtaining a result comparable with the performance of RF using manually crafted features. The scope of these experiments was limited to classes with more than 500 objects, including contact binaries, detached binaries, and three types of RR Lyrae, Long period variables (LPVs) and RS Canum Venaticorum variables (RS CVns). Their study involved experimenting with diverse CNN structures, background subtraction methods, and $\Delta m \Delta t$ binning approaches. They discovered that finer binning and effective background removal enhance classification efficacy.

Naul et al. (2018) proposed a recurrent autoencoder to improve understanding of periodic variable star embedding. The measurement error in observations was considered by weighting the reconstruction error in the loss function. The embedding was then used for classification purposes with an RF. Results show that the automatic features extracted by the autoencoder are competitive with engineered features from (Richards et al., 2011; Kim and Bailer-Jones, 2016) regarding classification error.

Aguirre et al. (2019) presented a CNN architecture for classifying variable stars using light curves from multiple astronomical surveys. The model uses a more straightforward representation of differences in time and magnitudes in consecutive observations, namely the $(\Delta t_i, \Delta m_i)_{i=1}^{L}$ representation. The CNN was evaluated using data from the Optical Gravitational Lensing Experiment survey III (OGLE; Udalski et al., 2008), VISTA Variables in the Vía Láctea ESO public (VVV; Minniti et al., 2010), and Convection, Rotation and planetary Transit (CoRoT; Auvergne et al., 2009) surveys, each characterised by unique filters, cadences, and sky coverage. It achieved classification ACC comparable to or superior to an RF classifier. The CNN architecture includes two convolutional layers for pattern extraction, followed by classification layers. A data augmentation technique was employed to balance the training data across classes. Capable of classifying both classes and subclasses, the model reached 85% ACC on the validation set. These findings underscore the potential of input representation and CNNs for efficient, large-scale classification of variable stars in future astronomical surveys.

Tsang and Schultz (2019) provided a model to learn light curve representations and classify objects with novel patterns. To achieve this, they trained an autoencoder and a classifier together. The model was trained using the All Sky Automated Survey(ASAS; Pojmanski and Maciejewski, 2005) data, including eight types of variable stars. The preprocessing and training approach was similar to that of Naul et al. (2018), using 200 observations to train the autoencoder, with Δtime and normalised magnitude as the input representation. The loss function was defined by $\mathcal{L}_{AE} + (\lambda\mathcal{L}_{GMM} + \mathcal{L}_{CE})$, incorporating the autoencoder loss, Gaussian mixture model loss, and cross-entropy loss, respectively.

These loss components compete during the training process. The Gaussian mixtures, learned in the latent space, aim to detect outliers, i.e., new variability patterns. Experiments with sequential and joint training strategies showed similar results, confirming the feasibility of training these modules separately.

Becker et al. (2020) proposed an end-to-end RNN approach for variable star classification using an input defined by differences in magnitude and time as Aguirre et al. (2019). The network is trained on three datasets: OGLE-III, em Gaia data release II (Brown et al., 2016, *Gaia* DR2;), and WISE. Classes with more than 500 objects were included. The results show that the network achieves classification ACC comparable to that of the RF classifier while being faster and more scalable. The network performs better than RF for fainter objects for less biased datasets like WISE, showing it can extract signals from noisy light curves.

Jamal and Bloom (2020) compared a set of neural network architectures to classify periodic variable stars, which require a well-estimated period. Four known layers were compared: long short-term memory (LSTM), gated recurrent unit (GRU), temporal convolutional neural networks (tCNN) and dilated tCNN. They studied two types of architectures: (i) a direct classification scheme, where information was encoded, and after that, this representation was used in a classifier, and (ii) a composite scheme that considers reconstruction and classification simultaneously. The same strategies were applied in a multi-band setting, where two architectures to merge bands were tested (before and after encoding). Direct classification architecture, which includes metadata (e.g., colour, period, amplitudes and averaged magnitudes), always improves ACC. Early stopping and dropout techniques were implemented to reduce the over-fitting; however, no comments about the data shift problem were given. Some sub-classes (e.g., RR Lyrae sub-classes) were classified erroneously even when metadata were incorporated; they recommended the injection of prior knowledge about the importance of each feature.

Bassi et al. (2021) introduced two DL methods for classifying variable stars: a two-dimensional CNN (2D CNN) and the one-dimensional CNN-Long Short-Term Memory

(1D CNN-LSTM). OGLE-III and CRTS catalogues were used for evaluation purposes, considering five classes from OGLE-III and seven classes from CRTS. Classes with more than ∼500 objects were considered in both surveys. The 2D CNN method, which requires preprocessing the same as Mahabal et al. (2017), shows high performance when applied to the OGLE dataset (f1 score 0.85), but its effectiveness decreases within the CRTS dataset (f1 score 0.54). On the other hand, the 1D CNN-LSTM model, which directly utilises the original light curves without any preliminary modification, obtains an f1 score that is somewhat lower than the 2D CNN (0.71 in OGLE and 0.49 in CRTS); it benefits from having fewer parameters and reduced computational demands. The authors propose that future research could focus on refining the hyperparameters of the 1D CNN-LSTM model. Overall, this research highlights the effectiveness of the 1D CNN-LSTM in classifying light curves without necessitating data preprocessing.

Zhang and Bloom (2021) proposed a new ANN architecture called cyclic-permutation invariant neural network. This novel approach was based on the representation of phased light curves through polar coordinates. Experiments using two implementations (temporal convolutional neural network and residual neural network) were conducted, outperforming non-invariant baseline models in three catalogues, namely, OGLE-III, MACHO and the All-Sky Automated Survey for Supernovae (ASAS-SN; Jayasinghe et al., 2019). A randomised and stratified split of the above catalogues (training, validation and testing sets) was applied to assess the generalisation capability.

Szklenár et al. (2022) focused on creating a multiple-input ANN to classify variable stars using their light curve images and physical parameters, namely, period and the Wesenheit index, which is calculated as $W_{\rm index} = I_{\rm band} - 1.55(V_{\rm band} - I_{\rm band})$. The network architecture combines an information flow of convolutional layers for processing light curve images, which is concatenated in dense layers with physical parameters. The image fed into the neural network is a visualisation of the phased light curve, which is transformed into a 1-bit (black and white) image with a size of 512 × 512 pixels. Data for this study are sourced from the OGLE-III survey. An oversampling approach is applied based

on Gaussian process regression to create synthetic light curves. The ANN performance is evaluated in two scenarios: one for classifying six primary types of variable stars and another for 16 sub types. Incorporating the period and the Wesenheit index raises the ACC range from 77%-99% to 89%-99%. Despite these improvements, the classification of anomalous Cepheids remains challenging due to imbalanced class problems.

Abdollahi et al. (2023) designed a hierarchical approach utilising deep CNNs for classifying variable stars. The proposed scheme operates in a two-tiered manner: initially determining the main type of a variable star, followed by sub-type classification. This approach is particularly effective in enhancing ACC for less-represented classes in the dataset. The methodology involves using light curves and the period of the stars as inputs for the ANNs. An essential step in data preprocessing includes folding light curves based on the period and their subsequent binning to achieve a uniform length. Light curves from OGLE-IV are used, considering seven classes and sixteen subclasses. The implemented CNN model demonstrates impressive performance, with an ACC of 98% in main type classification and 93% in sub-type classification, outperforming RNN models in ACC and training efficiency. The hierarchical structure of the model shows notable improvement in classifying less populous classes like Anomalous Cepheids and Type-II Cepheids, which are typically challenging.

Kang et al. (2023) introduced a novel DL model for classifying periodic variable stars, addressing the challenge of imbalanced datasets. Their approach is based on an ensemble augmentation strategy, which involves creating synthetic light curves through Gaussian processes to strengthen underrepresented classes. The researchers designed two ANN architectures: a multi-input RNN and a combined RNN-CNN structure. The multi-input RNN is fed with light curve information, period, and amplitude, whereas the RNN-CNN hybrid also incorporates light curve images into the CNN component. The input image is a single-channel, 128 × 128 pixel image that displays a phase-folded light curve. Comparative analysis revealed that the compound RNN-CNN model exhibited superior performance, achieving a macro F1 score of 0.75, outperforming the 0.71 scores achieved by

the multi-input RNN model. It is worth noting that even providing the period as input, the macro F1 score did not exceed 0.75.

To summarise, ANNs are becoming a crucial technology to speed up discoveries in variable stars; hence, incorporating new methods to ensure their generalisation capability is necessary.

1.4.4. Machine learning classifiers for RR Lyrae stars

Due to their relevance, some articles have developed machine-learning models for classifying certain classes of variable stars; here, we focus on RR Lyrae classification.

Based on their pulsation modes, RR Lyrae can be divided into ab-type (RRab), c-type (RRc), double-mode (RRd), and e-type (RRe). RRab stars pulsate in the fundamental radial mode, RRc stars pulsate in the first-overtone radial mode, RRd stars pulsate simultaneously in the fundamental and first-overtone modes, and RRe stars may pulsate in the second overtone (see, e.g., Catelan and Smith, 2015, and references therein).

Gran et al. (2016) designed a procedure to search for RR Lyrae stars considering two steps: first, a candidates generation phase, which is based on features analysis; and second, automatic classification using machine learning models. The first step considers a magnitude variability test and filtering based on the characteristic range for the estimated period. The machine learning models were based on feature engineering of light curves and were trained to classify RRab stars. This search focused on the Galactic bulge's outer part, as observed by the VVV ESO public survey (Minniti et al., 2010). More than 1,000 RRab stars were found using this procedure.

Elorrieta et al. (2016) compared a set of machine learning classifiers for RRab in VVV. These models were trained using 68 features. They conducted two approaches to assess the model performance: cross-validation and a test over two independent sets, achieving similar results. While these results are interesting, the shift level (in features space) between each independent set and the training set was not studied.

Sesar et al. (2017) proposed a gradient tree boosting classifier to identify RR Lyrae stars in multi-epoch and asynchronous multi-band photometric data from the Panoramic Survey Telescope and Rapid Response System (PanSTARRS; Kaiser et al., 2010). The training set considered 1.9 million objects and applied a three-stage classification scheme. The first classifier, which included 20 features, reduced the candidates from 1.5 million objects to 1,500. The second maintained ten features from classifier one and added 40 features from periodograms (periods and their powers), improving by 15.0% the purity of the extracted RR Lyrae sample, as compared with previous filtered stars. Finally, a multi-class (RRab, RRc, and non-RR Lyrae) classification using 70 features (multi-band and features used in previous stages) was applied to the remaining objects (910, with 95.0% completeness and 66.0% purity). They highlighted that this stage was highly time-consuming ($\sim$30 minutes per object); because of this problem, they needed to filter data in the previous steps. Additionally, for multi-band period estimation, a template fitting procedure was conducted; for this, physically informed modelling was applied to find the period, and characteristic ranges for each subclass were considered to constrain the period search during the estimation.

Dékány and Grebel (2020) designed a bidirectional long-short term memory recurrent neural network to separate RRab stars from other periodic variable stars in the VVV survey (Minniti et al., 2010). Unlike recent end-to-end deep ANNs approaches for classifying variable stars (Aguirre et al., 2019; Becker et al., 2020), they decided to use phased-folded light curves and provide the model: magnitudes, binned phase from period estimation (P), binned phase with $2P$, and P. The direct incorporation of period estimation had two reasons: an inexpensive computation in the studied survey since this has a low number of observations per object, and the relevance of this feature for separating RRab from other variable star types. They applied the same object selection strategy to minimise the shift between training and testing sets. The approach has a $\sim$99% precision and recall when applied to objects with a signal-to-noise ratio above six.

1.4.5. Generative models for variable stars

The generation of synthetic data is a prevalent method to enhance model quality. This data can be produced through simulation methods based on analytical models (Blanton and Roweis, 2007) or by applying domain knowledge to modify existing data, known as data augmentation. Simulation, influenced by theory-driven models, often requires substantial computational resources. Conversely, the domain of data augmentation has seen limited exploration in the literature (Castro et al., 2017; Aguirre et al., 2019; Kang et al., 2023). In this field, various constraints hinder these methods, particularly the sequential nature of the data. These methods face challenges in interpolating physical parameters and handling the irregular intervals between observations, presenting significant challenges to effective data enhancement.

A few publications have recently leveraged deep generative modelling to design procedures based on generative adversarial networks (GANs) and variational autoencoders (VAEs) to create synthetic light curves of periodic stars. These approaches offer a novel alternative to address the data shift problem discussed in this thesis.

García-Jara et al. (2022) proposed GANs for creating synthetic astronomical light curves to improve the classification of variable stars, addressing challenges like limited data sizes and class imbalance. They proposed a GAN-driven data augmentation strategy for conditional generation, considering the star type and physical characteristics of irregularly spaced time series. They introduced a new evaluation metric for selecting suitable GAN models and a resampling method designed to delay GAN overfitting during training. This resampling technique modifies the class distribution by controlling the probability of sampling from each class in a scheme without replacement, adaptable to highly imbalanced or balanced sets by adjusting a parameter γ. Additionally, they developed two data augmentation techniques that generate plausible time series while preserving the properties of the original data, though these did not surpass the GAN-based methods.

Martínez-Palomera et al. (2022) proposed a deep generative model, the physically enhanced latent space VAE (PELS-VAE), using a VAE to create synthetic light curves of periodic variable stars based on their labels and physical parameters. This model integrates a conditional VAE with alternative layers: temporal convolutional networks and RNNs (LSTM and GRU). Trained on OGLE-III dataset light curves and physical parameters from *Gaia* DR2, it aims for a more accurate latent space representation. The model generates unseen light curves by specifying physical parameters, mapping these to the latent space through regression, and then sampling and decoding the latent vector. The transformation of RR Lyrae light curves from saw-tooth to sinusoidal shapes with temperature increase showcases the model's effectiveness. This research presents an ad-hoc generative model that proficiently produces realistic light curves for periodic variable stars based on their physical parameters, offering new directions for data augmentation methodologies research.

Ding et al. (2024) proposed an unsupervised approach to classify contact binaries candidates; an autoencoder model is trained using synthetic light curves according to physical parameters, e.g., the mass ratio, the orbital inclination, the effective temperature of the primary star, the effective temperature of the secondary star. These parameters are provided to PHysics Of Eclipsing BinariEs software (Prša and Zwitter, 2005) to create synthetic samples. Once the autoencoder is trained, a sequence of criteria is created to detect potential contact binaries.

According to the reviewed articles, the generation of synthetic samples employing ML is a field that has been weakly explored, but its potential has been described clearly. Using a controllable model for generating samples conditioned to some physical parameters can reduce biases beyond the imbalanced class issue.

1.4.6. Bayesian data analysis in astronomy

In recent decades, several astronomical research articles have proposed applying a Bayesian analysis. For example, pioneering research was conducted by Gregory and Loredo (1992);

Saha and Williams (1994) on the parameter estimation of astrophysical models. The research field most heavily influenced by these developments has probably been that of cosmological parameter estimation (Christensen and Meyer, 1998; Christensen et al., 2001). Accordingly, Trotta (2008) provided a comprehensive review of Bayesian statistics emphasising cosmology. Sharma (2017) produced a literature review that focuses on the Monte Carlo Markov Chain (MCMC) for Bayesian analysis in astronomy, providing an extensive overview of several MCMC methods while also emphasising how astronomers have used Bayesian data analyses in the past and how such approaches should be used more commonly in the present. Furthermore, Sharma (2017) exemplified several basic concepts to model selection in a Bayesian approach. Subsequently, Hogg and Foreman-Mackey (2018) provided a pedagogical overview of MCMC in astronomical contexts and discussed its foundations, highlighting certain aspects to consider in order to avoid obtaining misleading results from applications of this otherwise powerful technique. Moreover, several papers have shown the advantages of the Bayesian model selection approach (Parviainen et al., 2013; Ruffio et al., 2018) in astrophysical model selection.

Weinberg (2013) presented a system to apply Bayesian statistics in astronomy, including methods for estimating the posterior distribution and managing the model selection. Furthermore, this scientific article provided a comprehensive introduction to Bayesian inference at both levels and included several case studies on the system in astrophysical models. Furthermore, Budavári et al. (2017) designed a method to decide whether observations correspond to particularly faint objects or noises from among data sets obtained in multiple epochs. Their marginal likelihoods are then compared. This approach exemplifies certain advantages of the Bayesian framework that stem from adding physical knowledge to hypothesis comparison.

Even though several papers have applied BMS to astronomy, according to our best knowledge, this is the first approach to adding physical information during the assessment process of machine learning classifiers for variable stars.

1.5. Publications

Many of the findings and content included in this work have already been published. In the following section, a list of the publications corresponding to each chapter is provided.

Informative Bayesian model selection for RR Lyrae star classifiers (Pérez-Galarce et al., 2021) is a journal paper published in Monthly Notices of the Royal Astronomical Society. It is focused on the first thesis objective and is presented in Chapter III.

Informative regularization for a multi-layer perceptron RR Lyrae classifier under data shift (Pérez-Galarce et al., 2023) is the second journal paper published in Astronomy and Computing, focusing on the second thesis objective and is presented in Chapter IV.

Finally, **A self-regulated convolutional neural network for classifying variable stars** was submitted to Monthly Notices of the Royal Astronomical Society covering the third objective and presented in Chapter V.

1.6. Outline

The document is organised as follows:

Chapter 2 provides the theoretical background for the methodologies proposed in this thesis, detailing the essential mathematical methods. It provides basic terminology for model selection and regularisation strategies. Furthermore, the chapter introduces the core machine learning models that support the proposed approaches.

In Chapter 3, the informative Bayesian model selection is presented. Section 3.2 outlines the proposed methodology for integrating simple deterministic rules into the model selection process. Section 3.3 describes the data used to evaluate the effectiveness of this approach. After that, Section 3.4 shows the results and discusses the impact of the methodology on improving model selection under data shift conditions.

Chapter 4 presents the methodology for incorporating human knowledge into ANN models, specifically focusing on the regularisation aspects. The shifted data for RR Lyrae classification is given in Section 4.4, detailing how the data shift problem is addressed in this context. Experiments comparing the approach undertaken with baseline alternatives are outlined in Section 4.5, demonstrating the advantages of the proposed method.

Chapter 5 shows a self-regulated CNN methodology. Section 5.6 expounds on the dataset, incorporating *Gaia* DR3 physical parameters and OGLE-III light curves to showcase the application of the methodology. Section 5.7 details our findings, emphasising how the self-regulation mechanism within the CNN enhances classification performance.

Finally, Chapter 6 provides a conclusion, summarising the main discoveries, discussing the implications of the findings, and suggesting avenues for future research lines, including potential enhancements to the methodologies.

2. CHAPTER II. PRELIMINARIES

2.1. Model selection

In the machine learning and statistical learning fields, the model assessment process is a central topic and one which is typically associated with three main tasks: (i) the evaluation of a population error using the training data error; (ii) the selection of the most suitable model among a set of alternatives; and (iii) the definition of a good set of hyper-parameters. This section summarises the traditional methods used to assess models.

2.1.1. Metrics for evaluating classifiers

There are several metrics to evaluate the performance of classification models, which have been originated from different fields such as statistical learning, information theory and data mining. Consequently, selecting one metric or a set thereof to assess our models can take time and effort. When selecting a metric, it is essential to consider the following well-known basic properties.

- Consistency: the size of the training data should not affect our metric.
- Occam's razor principle: we desire a metric that can identify whether a model has the optimal complexity.
- Comparison: It should allow us to compare non-nested models. A nested model is a model in which the features are a subset of those in another model, and both models share the same structure, e.g. both are logistic regressions.
- Reference: the metric must be independent of the validation strategy.
- Individuality: the metric should be able to measure any given object individually.

Below is a summary of the most frequently used metrics and validation strategies.

2.1.2. Metrics based on confusion matrix

The ACC is the most intuitive metric within this framework, which evaluates prediction quality based on the ratio of correct predictions over the total number of observations. This metric has two critical drawbacks: it cannot discriminate the type of error (false-positive and false-negative), and the majority class can easily dominate it.

Other measures can be obtained to analyse each type of error. For example, the *Recall* score, which represents the fraction of positive patterns that are correctly classified. Alternatively, the *Precision* score, which relates to the ratio between the positive objects that are correctly predicted and the total number of predicted objects in a positive class. To consider a balance between *Recall* and *Precision*, we can evaluate these two metrics in conjunction through the F_1-score. This metric is the harmonic mean between *Precision* and *Recall*; it is more robust than ACC when the dataset has imbalanced classes.

Although many variants exist in the literature, the metrics above are the most commonly used in this framework. A summary of these can be found in Sokolova and Lapalme (2009). Despite the large variety of metrics, several related limitations exist. First, we cannot directly compare the trade-off between data fitting and model complexity. Second, we must rely on validation strategies. Third, different confidence levels in predictions cannot be considered, as the metrics are based on hard classification; that is, given a threshold (e.g., 0.5), we assume a binary decision regarding the predicted class.

2.1.3. Bayesian model selection

A robust alternative for selecting models is the marginal likelihood, which is denoted by $p(\mathcal{D}|m)$, where m represents a model and $\mathcal{D}$ is the data. It appears in the first level of inference in the Bayesian framework,

$$p(\theta|\mathcal{D}, m) = \frac{p(\mathcal{D}|\theta, m)p(\theta|m)}{p(\mathcal{D}|m)}, \tag{2.1}$$

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\theta, m)p(\theta|m)d\theta. \tag{2.2}$$

We use the following traditional notation: let $p(\theta|\mathcal{D}, m)$ be the posterior distribution of the parameter given the data and a model; let $p(\mathcal{D}|\theta, m)$ denote the likelihood function; let $p(\theta|m)$ represent the prior distribution over the parameters and finally, let $p(\mathcal{D}|m)$ be the marginal likelihood.

The marginal likelihood like a model selector was analysed in-depth by MacKay (1992) and, subsequently, the links with Occam's razor principle were emphasised in Rasmussen and Ghahramani (2001); Murray and Ghahramani (2005); Ghahramani (2013). The idea of using the marginal likelihood in model assessment comes from the second level of inference,

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m\in\mathcal{M}} p(m, \mathcal{D})}, \tag{2.3}$$

where the Bayes' theorem is used to estimate the model probability given a dataset $p(\mathcal{D}|m)$

Estimating $p(m|\mathcal{D})$ is intractable since we cannot enumerate all possible models. However, the same criteria used in the first level of inference can be applied, avoiding the denominator estimation (constant). In this way, we can estimate the model posterior by $p(m|\mathcal{D}) \propto p(\mathcal{D}|m)p(m)$. Finally, if we assume a non-informative prior for the models, $p(m|\mathcal{D})$ is proportional to the marginal likelihood. For this reason, we use the marginal likelihood to select the most appropriate model.

Even though the marginal likelihood automatically embodies all the desired properties of good metrics, it can not automatically manage the biases in the training data, and its estimation is a computational challenge in high-dimensional data (see integral in equation 2.2). To address this challenge, we can estimate the marginal likelihood by interpreting it as an expected value and then performing Monte Carlo (MC) estimation according to the following equation:

$$p(\mathcal{D}|m) = \mathbb{E}_\theta \left[p(\mathcal{D}|\theta, m)\right], \tag{2.4}$$

$$\frac{1}{S}\sum_{s=1}^{N} p(\mathcal{D} \mid \theta_s, m), \theta_s \sim p(\theta). \tag{2.5}$$

This simple approach only performs well if the prior and likelihood have a similar shape and are overlapped strongly. In different scenarios, misleading samples can be generated in low-valued areas of the likelihood function. Subsequently, a few samples with high values in the likelihood function dominate the estimator, and this could produce a high variance in the estimation procedure.

Due to these difficulties, the majority of research into BMS avoids MC sampling methods by applying approximations such as the Laplace approximation and penalised likelihoods (Schwarz et al., 1978; Watanabe, 2013) or they resort to MC methods based on posterior samples (Neal, 2001; Raftery et al., 2006; Overstall and Forster, 2010).

We resort to bridge sampling, which offers a variant for the last type of strategy. The bridge sampling approach begins with this identity,

$$1 = \frac{\int p(\mathcal{D}|\theta)p(\theta)h(\theta)g(\theta)d\theta}{\int p(\mathcal{D}|\theta)p(\theta)h(\theta)g(\theta)d\theta}, \tag{2.6}$$

where $g(\theta)$ is the proposal distribution. Subsequently, it is multiplied by the marginal likelihood on both sides. Then, we obtain the following equation:

$$p(\mathcal{D}) = \frac{\int p(\mathcal{D}|\theta)p(\theta)h(\theta)g(\theta)d\theta}{\int \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}h(\theta)g(\theta)d\theta}. \tag{2.7}$$

Note that the posterior distribution appears on the right side of the denominator. After that, through

$$p(\mathcal{D}) = \frac{\int p(\mathcal{D}|\theta)p(\theta)h(\theta)}{\int h(\theta)g(\theta)}\frac{g(\theta)d\theta}{p(\theta|\mathcal{D})d\theta}, \tag{2.8}$$

it is separated into two ratios, and after that, we can obtain the expected values in the denominator and numerator as follows:

$$p(\mathcal{D}) = \frac{\mathbb{E}_{g(\theta)}\left[p(\mathcal{D}|\theta)p(\theta)h(\theta)\right]}{\mathbb{E}_{p(\theta|\mathcal{D})}\left[h(\theta)g(\theta)\right]}. \tag{2.9}$$

Finally, we use the definition of optimal bridge function provided by (Meng and Wong, 1996), which is presented in the next equation:

$$h(\theta) = C\frac{1}{s_1 p(\mathcal{D}|\theta)p(\theta) + s_2 p(\mathcal{D})g(\theta)}. \tag{2.10}$$

Since this expression depends on the marginal likelihood, to tackle this recurrence, it is applied the iterative scheme presented below:

$$\hat{p}(\mathcal{D})^{t+1} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2}\frac{p(\mathcal{D}|\theta_i)p(\theta_i)}{s_1 p(\mathcal{D}|\theta_i)p(\theta_i)+s_2\hat{p}(\mathcal{D})^t g_(\theta_i)}}{\frac{1}{N_1}\sum_{j=1}^{N_1}\frac{g(\theta_j)}{s_1 p(\mathcal{D}|\theta_j)p(\theta_j)+s_2\hat{p}(\mathcal{D})^t g_(\theta_j)}}. \tag{2.11}$$

2.1.4. Validation strategies

CV is the most common family of methods for estimating metrics. This section presents an overview and summarises some drawbacks of the three most common CV-based methods. Firstly, *Hold-out*, which is the most basic approach. Therein, two data sets are generated; one is used to train the model, and the other is used to evaluate its quality. This approach depends heavily on one dataset, and for that, it is a good option only when there are large quantities of data or when running time limitations exist. Moreover, hold-out can generate a pessimistic estimator (Lendasse et al., 2003) in small datasets. Secondly, *k-fold*, which is the most commonly used variant of CV, considers splitting the data $\mathcal{D}^S$ into smaller chunks $\mathcal{D}_1^S, \mathcal{D}_2^S, ...\mathcal{D}_K^S$ with the same size. We train using $\mathcal{D}^S \setminus \mathcal{D}_k^S$ chunks and evaluate using the free chunk $\mathcal{D}_k^S$. The number of folds provides the bias-variance trade-off; a small number of folds may reduce the bias but also may increase the variance. Lastly, *Leave-One-Out* uses each data point as a chunk, and for each object, a model is trained to leave only this object out. Although its variance can be larger, it provides an unbiased estimator, but its running time can be prohibitive (Rao et al., 2008).

Arlot et al. (2010) presents a survey of CV procedures for model selection. Despite the effort to develop variants, these ideas consist of a fundamental assumption. They consider that we are working with representative training data. It means that $\mathcal{D}^{\mathcal{S}}$ have the same

probability distribution that data beyond the labelled objects $\mathcal{D}^{\mathcal{T}}$. However, in astronomy, we cannot generate such representative training data.

In an attempt to overcome this challenge, Sugiyama et al. (2007) proposed a CV variant to tackle the biases above (also known as *data shift* problem) through an importance-weighted CV ($IWCV$) approach. The $IWCV$ weighs each observation i in the evaluation metric with the density ratio $p(x_i)_{\text{test}}/p(x_i)_{train}$. Note that IWCV is proposed to address a type of *data shift*, which is named as a covariate shift (bias in features), here, $p(x_i)_{\text{train}} \neq p(x_i)_{\text{test}}$, but $p(y|x)_{\text{train}} = p(y|x)_{\text{test}}$. If one needs to address scenarios involving bias in labels (Target Shift), where $p(y)_{\text{train}} \neq p(y)_{\text{test}}$, but $p(x|y)_{\text{train}} = p(x|y)_{\text{test}}$, an adaptation of the density ratio by $p(y)_{\text{train}}/p(y)_{\text{test}}$ becomes necessary. While this is a clever approach, it presupposes existing knowledge of the probability density functions for $\mathcal{D}^S$ and $\mathcal{D}^T$, which can be intractable in high dimensions.

A further commonly used method is bootstrap (Efron and Tibshirani, 1997); it uses sampling with replacement, in which N samples are selected in each of the k iterations; in this approach, a sample could appear more than once in each iteration, and the testing contains instances that have never previously been seen. Thus, each step selects an instance with probability $1/N$. Although the traditional bootstrap has interesting statistical properties, it fails to select classifiers in a machine-learning context because it favours overfitting classifiers (Kohavi et al., 1995).

An interesting variant is found when bootstrap is analysed from a Bayesian perspective (Rubin, 1981). A traditional bootstrap can be understood by modelling the probability of drawing a specific observation such as a categorical distribution, *Cat(π)*, where the vector $\pi = (\pi_1, \pi_2, ..., \pi_N)$, is the probability of drawing each object ($\sum_i^N \pi_i = 1$). In a traditional bootstrap we have $\pi_1 = \pi_2 = \pi_k = \pi_N = 1/N$. In a Bayesian view, π draws from a Dirichlet, *Dir(α)*, where for example, the expected proportion for π_1 is based on the priors $\alpha_1/\sum_{i=1}^N \alpha_i$. However, generating informative priors $Dir(\alpha)$ can pose a significant challenge.

2.2. Prior knowledge injection into neural networks

The injection of human knowledge in ANNs (and into machine learning models in general) is one of the longest-standing challenges in artificial intelligence. Deng et al. (2020), Von Rueden et al. (2021), and Borghesi et al. (2020) provide comprehensive literature reviews about how human knowledge has been integrated into machine learning models. Most of these approaches have used inductive biases, i.e., a prioritisation of solutions independent of the observed data e.g., regularisers, prior distributions, or data augmentation (Battaglia et al., 2018). The main lines of research undertaken in this area are presented below:

Data augmentation: Using human knowledge to generate synthetic data is a common strategy for data quality improvement. This data can be generated by using: simulation techniques from analytical models (Blanton and Roweis, 2007); applying simulation based on machine learning model (Sravan et al., 2020); or using domain knowledge to modify the available data. To illustrate, in the domain of images, knowledge about image invariance is the basis for several standard alternatives of data augmentation, such as flipping, cropping, scaling and translation (Shorten and Khoshgoftaar, 2019). Regarding light curves, few scientific articles have focused on data augmentation (Castro et al., 2017; Aguirre et al., 2019). In this domain, some conditions limit these approaches, such as the temporal dependency of observations; often, light curves must be modelled as a multivariate time series (magnitude observed in more than one band), and the time between observations is irregular. Lastly, to avoid more biases due to the augmented data, knowing the less-represented patterns in light curves that can be challenging to detect in advance is necessary.

Bayesian modelling : This approach is intuitive for adding human knowledge to machine learning models. The prior distributions offer a direct path to provide expert knowledge. However, literature on informative prior distributions for complex models (e.g., MLP or

AE) is scarce (Hanson et al., 2014; Fortuin, 2022) and obtaining the true posterior is highly time-consuming.

Knowledge base: Typically represented by graphs, knowledge bases provide a structure that offers a simple alternative for storing knowledge. Examples include Freebase (Bollacker et al., 2007), DBpedia (Auer et al., 2007), and YAGO (Suchanek et al., 2007). Knowledge graphs consider entities represented by nodes and relations represented by arcs. The knowledge contained therein is declarative as it uses symbolic language and has been used to generate new data and provide explainability. Kafle et al. (2020) review how knowledge bases are incorporated into ANNs. In broad terms, there are two well-known drawbacks to this structured knowledge base, and these relate to completeness and compatibility. The completeness issue comes from the fact that there is no graph containing all the available information. Compiling this knowledge base could be hugely challenging in specific scientific disciplines, such as astronomy. The compatibility problem relates to the design decisions of each graph, whereby each decision may consider different relations among entities.

Feature space: Traditional mechanisms for feature selection are based on compressing original space or selecting an optimal subset of features. Such methods are purely data-driven, and prior knowledge is not included. Atzmueller and Sternberg (2017) propose generating an additional set of features using a domain-specific knowledge graph for combining pairs of features. They suggest that this knowledge graph successfully manages the level of interaction among features during the knowledge-based feature generation.

Hypothesis space: The hypothesis space considers the weights and architecture (Borghesi et al., 2020). Direct knowledge injection on weights can be performed using constraints or penalisations during optimisation. The knowledge injection on architecture is incorporated using ad-hoc layers. For example, convolutional layers use knowledge about the correlation of neighbour pixels, recurrent layers use temporal relations knowledge, and graph neural networks have been designed to use graph-expressible prior information.

High-level knowledge: This approach incorporates physical or logic rules into machine learning models. Differential equations can represent the physical rules, while logic rules are supported by first-order logic. Both knowledge classes can be incorporated as a new loss function. Raissi et al. (2019) proposed physically informed ANNs to manage the learning process under a small data set. This type of neural network is constrained to certain symmetries, invariances, or conservation principles whose origin comes from the physical laws and can be represented by non-linear partial differential equations. In this setting, two functions are included: first, a loss minimising the initial and boundary data, and second, a loss in learning the structure imposed by the non-linear partial differential equation.

Based on these contributions, the second proposal in this thesis seeks to emphasise the mitigation of the data shift problem in MLPs by combining: a knowledge representation using highly informative rules; sparse training cases generation (signals) based on this knowledge representation; and prioritisation of solutions (weights) coherently with deterministic rules through regularisation and an ad-hoc training procedure.

It should be noted that a significant component of the research undertaken for this contribution is the regularisation scheme; therefore, this topic will be reviewed in the next section.

2.3. Regularisation strategies

This section provides a review of regularisation strategies. First, Section 2.3.1 discusses the classic non-informative approaches. Subsequently, Section 2.3.2 examines proposals for incorporating human knowledge into regularisers.

2.3.1. Non-informative regularizers

Regularisation in ANNs includes strategies focused on improving generalisation capability, i.e., reducing the over-fitting issue on over-parameterised ANNs. Goodfellow et

al. (2016) reviews the main techniques within the regularisation framework, such as norm penalties, dataset augmentation, noise robustness, early stopping, parameter tying/sharing, sparse representation and dropout. Each method has been applied intensively in recent years; however, a gap remains concerning injecting human knowledge to mitigate overfitting. It should be noted that overfitting on shifted data is intractable if we only rely on the data itself, as no reliable datasets are available to learn patterns that generalize well or to validate performance. Thus, knowledge incorporation beyond the training data set is critical.

Neyshabur et al. (2014) discussed the inductive bias, which can be understood as a capacity controller for improving generalisation. This inductive bias must consider appropriate flexibility to fit the data accurately. Indeed, Neyshabur et al. (2014) used the number of hidden units for experimental purposes by running a comparison with the matrix factorisation strategy. They studied the impact on training/testing using a dimensional control (hidden size) versus norm control (weights). The authors comment that their optimisation strategy incorporates a regularisation bias towards "low complexity" global minima.

Leimkuhler et al. (2020) proposed a constrained approach to regularise ANNs, which the authors apply into a stochastic gradient Langevin optimisation algorithm. These algebraic constraints control the weight values, and with this in mind, the authors propose the following three constraints: circle constraints, sphere constraints and orthogonality constraints. The generalisation capability is improved when constrained training and unconstrained approaches are compared to training MLPs on different image sets for classification tasks. Despite achieving good performance, there is no human knowledge injection using these constraints.

A general strategy to induce regularisation on machine learning models employing a penalised component can be defined as follows:

$$g(\beta) = f(\beta) + \lambda\Omega(\beta), \tag{2.12}$$

where β represents the model parameters; $f(\beta)$ is a score function, assessing the prediction quality; and $\Omega(\beta)$ is the penalised component, which is a function of β, typically, a norm. λ is a tuneable hyperparameter that controls the relative importance of the penalised component with respect to the score function during the optimisation process.

A well-known penalised approach, which is used from simple regression to complex deep neural networks, is the least absolute shrinkage and selection operator (Lasso or l_1-norm; Tibshirani, 1996). This regularisation corresponds to norm-based regularisers. In a classification model using l_1-norm, $f(\beta)$ can be a cross-entropy function, β represents the model parameters that are controlled/penalised by the sum of absolute values of weights $||\beta||_1$, which is represented by $\Omega(\beta)$ in equation (2.12). A cross-polytope defines the feasible solutions space for the weights from a geometric interpretation. This regularisation has a Bayesian interpretation considering a Laplacian prior over the weights. Such a method lacks human knowledge to define the space of feasible solutions, and only the relationship between complexity and weight values is used. This relationship comes from the fact that large weight values increase the output variance; then, according to the bias-variance error decomposition, a large variance increases the generalisation error (Geman et al., 1992; Svozil et al., 1997).

Similarly, ridge regression constrains the weights by the l_2-norm. Here, the geometric interpretation is associated with a hyper-sphere, and the Bayesian perspective is linked to a Gaussian prior (Hoerl and Kennard, 2000). Unlike Lasso, the ridge model does not provide a straightforward model interpretation because the parameters only reduce their absolute weight values while remaining non-zero. By focusing on interpretability and due to the poor performance of Lasso regression when the number of features exceeds the number of observations ($p \gg n$ problem), Zou and Hastie (2005) proposed the elastic net. The latter can be considered a convex combination between the l_1-norm and l_2-norm. Unlike the l_1-norm, this generalisation can manage the grouping problem, which occurs when a group of features has a high pairwise correlation. Conversely, the l_1-norm selects only one variable and places no importance on which one it is (Zou and Hastie, 2005).

More recently, Kim et al. (2021) have proposed a regularisation approach to control the well-known imbalance problem. This regulariser considers two components of the output model distribution: a mean divergence regularisation,

$$\Omega(\beta) = \frac{\lambda}{2}\left[\frac{1}{2} - E_x\left[f_\beta\left(x\right)\right]\right]^2, \tag{2.13}$$

where f_β function represents the classifier; and a variance divergence regularisation based on KL divergence,

$$\Omega(\beta) = \frac{\lambda}{2}\left[D_{KL}(Z_+||Z_-) + D_{KL}(Z_-||Z_+)\right], \tag{2.14}$$

where Z_+ and Z_- represent positive and negative training cases, whose distributions are assumed to be $\mathcal{N}(\mu_1, \sigma_1)$ and $\mathcal{N}(\mu_2, \sigma_2)$, respectively.

These regularisation methods outperformed baseline models, mitigating the imbalance problem in classifying sentences and images. They also highlighted the challenge posed by selecting the λ hyperparameter. Hence, new methods avoiding this hyperparameter search are crucial.

To summarise, several ideas have been proposed to regularise weight-based models, including sparseness promoting, weight controlling and ad-hoc loss functions for the imbalance problem. However, these approaches have been focused on improving the generalisation of representative data. Therefore, when the data are shifted, these approaches become futile. Therefore, combining such methods with human knowledge is necessary to improve the generalisation capability in this complex setting.

2.3.2. Informative regularizers

Only a few papers have incorporated knowledge into regularisation components, and this section provides examples of exciting approaches in that direction.

Applying the aforementioned non-informative approaches as regularisers would lead to an unstable model (i.e., one vulnerable to resampling). To mitigate this drawback,

Baldassarre et al. (2012) compared the l_1-norm, elastic net, sparse Laplacian and total variation methods with specific alternatives that exploit spatial information. Chambolle (2004) proposed the total variation method, which uses an additional component to the penalisation term:

$$\Omega(\beta) = ||\nabla\beta||_1 + ||\beta||_1, \tag{2.15}$$

where $\nabla\beta$ represents the difference between neighbours (e.g., voxels or pixels). For example, the difference in the first dimension can be computed as $(\nabla\beta)^1_{ijk} = \beta_{ijk} - \beta_{(i+1)jk}$. This regularisation of the weights promotes similar behaviour in neighbouring voxels in addition to sparsity. Sparse Laplacian offers a soft alternative in which the constant requirements are relaxed by the incorporation of an additional tunable parameter α,

$$\Omega(\beta) = \lambda(1-\alpha) \sum_{(i,j)\in\xi_G} (\beta_i - \beta_j)^2 + \lambda\alpha||\beta||_1, \tag{2.16}$$

with ξ_G being the set of edges over the graph of connections G. If $\lambda_1 = \alpha\lambda$ and $\lambda_G = \lambda(1-\alpha)$, sparse Laplacian is equivalent to graph Laplacian elastic net (graph-net). Based on the spatial informative regulariser, Jenatton et al. (2012) proposed a hierarchical and spatial informative penalisation. In this approach, the neighbourhood ξ_G can be represented using a tree defined from a spatially constrained agglomerative clustering. This spatial and hierarchical regulariser provides enhanced interpretability and improves accuracy in predictions. This spatial informative regulariser type has also been used in SVM models considering embedded features selection (Watanabe et al., 2014).

In summary, limited contributions have been made to improve the generalisation capability of using informative regularisers. The knowledge added has been primarily focused on spatial information and is insufficient to mitigate the data shift problem in a time-series domain. Therefore, regularisation approaches need to be boosted by injecting high-level knowledge.

2.4. Multi-layer perceptron notation

Let $\mathcal{N}$ be an artificial neural network with a multi layer perceptron (MLP) architecture which defines a mapping between two Euclidean spaces $(\mathbb{R}^{n_1}, \mathbb{R}^{n_L})$. This mapping is obtained employing a sequence of operators in each layer i and neuron j, consisting of an affine transformation,

$$z_{ij}(x_k) = \sum_{k=1}^{n_{i-1}} w_{jk}^{i} x_k - \theta_j^i, \tag{2.17}$$

followed by a non-linear function (activation). Multiple alternatives are available to define the activation functions, the most common being the sigmoid activation function,

$$g_{ij}(x) = \sigma\left(z_{ij}(x)\right), \tag{2.18}$$

$$\sigma(z) = \frac{1}{1+e^{-z}}, \tag{2.19}$$

and the rectified linear unit (ReLu) activation function,

$$g_{ij}(x) = \max\{0, z_{ij}(x)\}. \tag{2.20}$$

In this mathematical model, $g_{ij}(x)$ represents the activation degree in the neuron j of layer i. $n_1, n_2, ..., n_L$ are the number of neurons in each layer and L defines the number of layers. x_k is the input value in dimension k. The dimensionality of the first layer input is $\mathbb{D} \in \mathbb{R}^{n \times m}$, where n is the number of features and m the number of objects in the training set (e.g., variable stars). Moreover, each object is defined by the tuple (d, l), where $d \in \mathbb{D}, l \in \mathbb{L}$. To train this model, optimisation techniques, e.g. mini-batch gradient descent, are applied to predict a label $l \in \mathbb{L}$ for each object, minimising the error in the final mapping with respect to the label mentioned above. The optimisation process, typically conducted using a back-propagation variant, searches for a good combination of solutions for each W^i and θ^i. $W^i \in \mathbb{R}^{n_i \times n_i}$ denotes a weight matrix containing each learnable parameter w_{jk}^i (weights) and θ^i matrix includes decision variables θ^i (biases).

For binary classification, $\mathcal{N}$ has a one-dimensional output layer in which the activation degree represents the probability of success for this training case.

3. CHAPTER III. INFORMATIVE BAYESIAN MODEL SELECTION FOR RR LYRAE STAR CLASSIFIERS

3.1. Method overview

The Bayesian model selection framework relies on the marginal likelihood (also known as Bayesian evidence), which is the likelihood function weighted by a prior distribution over the range of values for its parameters. In other words, the marginal likelihood represents the expected probability of the data given the parameters. However, if additional information is not incorporated into these prior distributions, even this powerful and robust metric may fail to assess models accurately when training data are biased. To address these biases, we propose a strategy that leverages expert knowledge by incorporating informative priors in the marginal likelihood estimation for RR Lyrae star classifiers.

Our methodology is divided into three stages. First, we present a method to represent prior knowledge using deterministic rules based on physically meaningful features, such as period and amplitude. Using this representation, we estimate a likelihood function, which serves as the prior distribution in subsequent steps. In the second stage, we generate posterior samples using these informative priors. This approach ensures that high-value regions in both the likelihood function and prior distribution are prioritized while also allowing the integration of astronomical knowledge through the influence of priors on the posterior distribution. Finally, in the third phase, we estimate the marginal likelihood using an approximate sampling method.

3.2. Method description

This section provides a comprehensive description of our method for adding human knowledge to the assessment of variable star classifiers. The methodology assumes that we have a set of models $\{m_1, m_2, .., m_i, ..m_n\} \in \mathcal{M}$ and a biased set of objects (variable

stars) to train them. Our goal is to sort these models according to their performance in a shifted data (testing set).

The method can be divided into three main steps. Section 3.2.1 focuses on obtaining priors from deterministic rules. Section 3.2.2 considers the generation of posterior samples running an MCMC algorithm. Section 3.2.3 presents the mechanism for adding the informative posterior samples to the marginal likelihood estimation procedure.

Figure 3.1 shows a diagram of our method, in which the output for each step is highlighted. The final output is a ranking of models based on an informative estimation of the marginal likelihood. We propose mitigating biases by means of this informative marginal likelihood.

3.2.1. Obtaining informative priors

In the Bayesian framework, informative priors offer a great opportunity to add expert knowledge to machine learning models; however, the majority of Bayesian proposals use non-informative priors based on the likelihood function (Gelman et al., 2017). It can be controversial (Gelman et al., 2008; Golchi, 2019), and it is valid for both levels of inference; the first level, when we do inference on parameters; and in the second level, when we do inference on models. For some models, the addition of human knowledge can be less complex since it can be transferred from the space of features to the space of parameters directly; this is the case of Bayesian GMM or Bayesian naive Bayes. However, in models like Bayesian logistic regressions or Bayesian neural networks, it is not direct.

Regarding BLR, proposing informative priors can be a great challenge (Hanson et al., 2014). Despite that, some alternatives have been proposed to add expert knowledge when it is available. Gelman et al. (2008) proposed weakly informative priors (Cauchy prior) that are also useful to solve the complete separation problem (Zorn, 2005). Hanson et al. (2014) provided a Gaussian g-priors scheme. This approach is suitable if there is information about the probability of each class. In spite of these proposals, according to

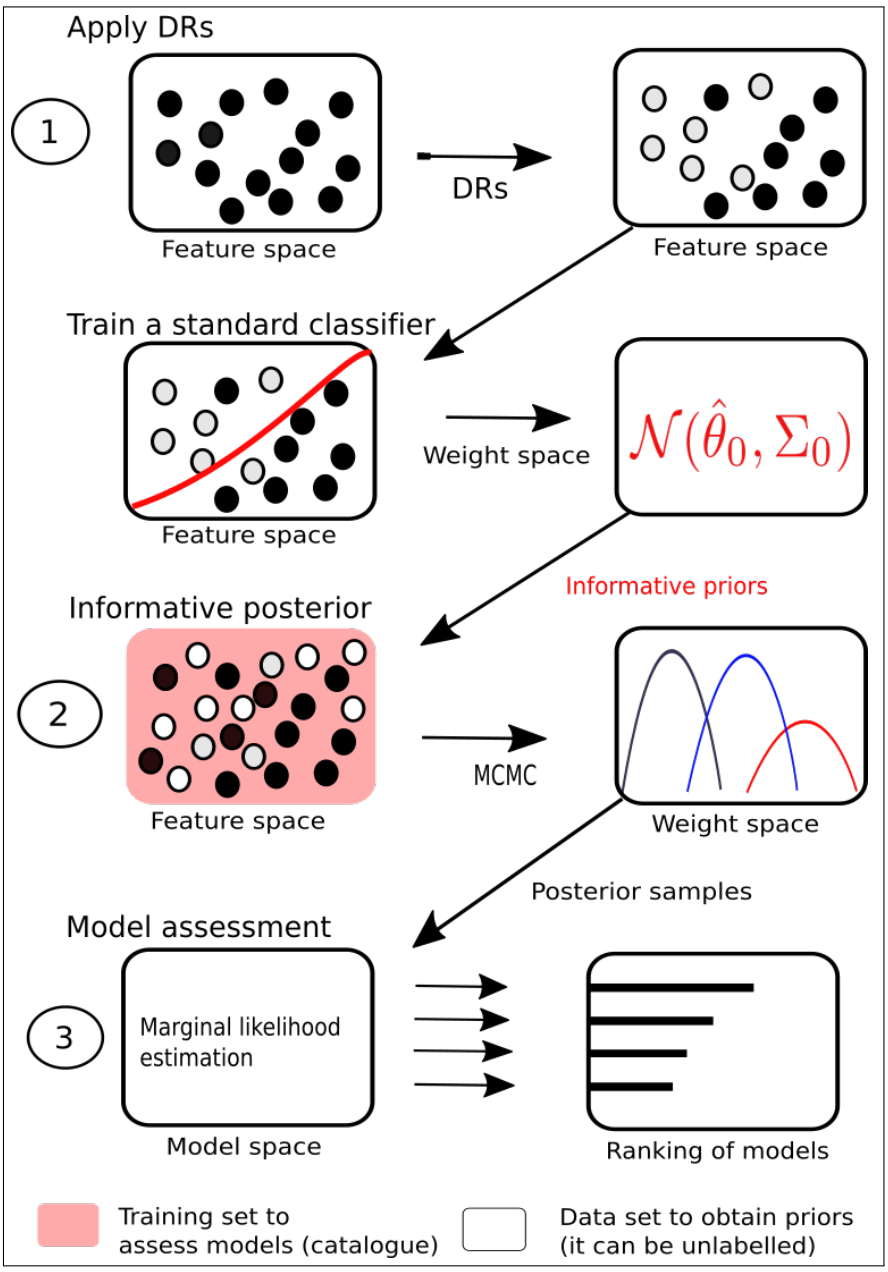


Figure 3.1. **Method overview of the informative Bayesian model selection.** 1) We first modify a variable star data set, assigning labels according to DRs. Then, we train a standard logistic regression to obtain its weights. 2) After that, the mean and variance of those weights are used in an MCMC frame to generate posterior samples. 3) The marginal likelihood is estimated by using these informative posterior samples. Lastly, considering the estimated marginal likelihood for each model, we generate a ranking of them.

our best knowledge, information about the relationship between classes and features can not easily be incorporated. To face this challenge, we propose the next methodology to obtain informative Gaussian priors for Bayesian logistic regression.

We propose to obtain astronomical knowledge through DRs. DRs can be used to filter celestial objects without resorting to machine learning methods. The DRs are based on physical features such as period, mean magnitude, and amplitude. To design these rules for RR Lyrae Stars, we can make use of literature in the field about features that may be particularly relevant in characterising this class of variable stars.

DRs can be understood as a relationship between an antecedent (if) and a consequent (then). To define a rule, we use a standard notation, $A \Rightarrow B$, where A represents a physical condition (antecedent) and B represents a class of variable stars (consequent). Certain examples of DRs for pulsating stars include:

- (period $\in [0.2, 1.0]$ days) $\Rightarrow$ RR Lyrae
- (amplitude $\in [0.3 - 1.2]$ in V-band) $\Rightarrow$ RR Lyrae
- (amplitude $\in [0.2 - 0.8]$ I-band) $\Rightarrow$ RR Lyrae
- (period $\in [1, 100]$ days) $\Rightarrow$ Classical Cepheid
- (period $\in [0.75, 30]$ days) $\Rightarrow$ Cepheid type II
- (period $\in [0.5, 8.0]$ hours) $\Rightarrow$ Dwarf Cepheids

It is important to highlight that these rules, intuitively, should be expressed as RR Lyrae $\Rightarrow$ (period $\in [0.2, 1.0]$ days). However, we use the other direction to manage the knowledge to label objects. Note that some physical condition can be valid for more than one variable star class, however when applying a chain of several DRs, this drawback is reduced. Despite that, we recommend mitigating this possible overlap by using not only various DRs but also DRs based on invariant features, e.g., period and amplitude (Catelan and Smith, 2015). Once we have obtained the DRs, we propose the application of Algorithm 3.2.1 to obtain informative priors. The algorithm is proposed to identify priors in a binary classification scheme; thus, we must use a set of rules for each class of variable stars.

The priors $\hat{\theta}$ are generated by fitting a standard (non-Bayesian) logistic regression. The training data with which to fit this model becomes critical at this stage. It is possible to use an entire survey, a subset of a survey, or even a modified survey (data augmentation, downsampling or oversampling). It is worth noting that we not only desire to label high-probable objects, but also we find to incorporate knowledge of relevant features for this class of variable stars.

This method allows astronomical knowledge to be transferred from the space of physical features to the space of model parameters through those collected deterministic rules of physical features for RR Lyrae variable stars. In particular, we define the mean estimator vector, $\hat{\theta}$, and the variance estimators, Var($\hat{\theta}$), for a normal prior. Var($\hat{\theta}$) is defined by the diagonal of the inverse Fisher information $\mathbf{I}(\theta)$. To avoid very small values for the estimated prior of variance Var($\hat{\theta}$), we add a small constant ϵ (for example, ϵ =0.1) after applying Algorithm 3.2.1.

Algorithm 3.2.1 Prior knowledge from DRs

Input: Data $\mathcal{D}_{\mathcal{X}}$, classifier m, DRs Rules
Output: weights for the classifier m, β
  $\mathcal{D}_{\mathcal{Y}} = \mathbf{1}$
  for $r \in$ Rules **do**
    for $d \in \mathcal{D}_{\mathcal{X}}$ **do**
      state$\leftarrow$ r.applyDR(d)
      if state$==$ False **then**
        $\mathcal{D}_{\mathcal{Y}}[d]$ = 0
      end if
    end for
  end for
  $\hat{\theta} \leftarrow$ m.fit($\mathcal{D}_{\mathcal{X}}, \mathcal{D}_{\mathcal{Y}}$)
  Var($\hat{\theta}$) $\leftarrow$ diag($\mathbf{I}(\theta)^{-1}$)
  return $\hat{\theta}$, Var($\hat{\theta}$)

3.2.2. Posterior samples generation

Our path for transferring human knowledge is by means of posterior samples since these contain both prior knowledge and data information. In this step, for each model we construct a Bayesian logistic regression in which priors come from the previous step.

To estimate the posterior $p(\theta|\mathcal{D}_{\mathcal{X}}, m)$, we propose the use of standard MCMC techniques, such as Metropolis-Hastings or Hamiltonian Monte Carlo algorithms, which are classic alternatives in the MCMC approach. Moreover, we use the Gelman-Rubin test to validate the sample convergence in each dimension (Gelman and Rubin, 1992). Lastly, to manage imbalanced classes, we downsample the datasets. This step is time-consuming; hence, more efficient sampling strategies could speed up our strategy. Variational inference approaches are not applied since the samples from those techniques are biased and our approach is based on good (precise and unbiased) samples (Blei et al., 2017).

3.2.3. Informative marginal likelihood estimation

The marginal likelihood has been widely studied to compare and select machine learning models, despite the fact that, its estimation represents a significant computational challenge. Comprehensive references for the study of estimation methods can be found in (Gronau et al., 2017; Wang et al., 2018).

We propose addressing an informative estimation of the marginal likelihood by using a bridge sampling approach (Overstall and Forster, 2010; Gronau et al., 2017); unlike standard Monte Carlo estimators (importance sampling or harmonic mean estimator), bridge sampling allows us to avoid dealing with typical constraints of standard Monte Carlo methods in relation to the shape of the proposal probability distributions. Indeed, this method has suitable properties in our context, mainly due to the following reasons: (i) it does not waste resources by generating samples in low-value zones, and (ii) it allows us to incorporate astronomical knowledge in order to reduce the impact of biases in the training data. The bridge sampling estimator is based on a ratio of two expected values as follows:

$$p(\mathcal{D}) = \frac{\mathbb{E}_{g(\theta)}\left[p(\mathcal{D}|\theta)p(\theta)h(\theta)\right]}{\mathbb{E}_{p(\theta|\mathcal{D})}\left[h(\theta)g(\theta)\right]}. \tag{3.1}$$

To estimate $\mathbb{E}_{g(\theta)}\left[p(\mathcal{D}|\theta)p(\theta)h(\theta)\right]$ we use samples from a proposal distribution, $g(\theta)$, and to estimate $\mathbb{E}_{p(\theta|\mathcal{D})}\left[h(\theta)g(\theta)\right]$ we need posteriors samples, $p(\theta|\mathcal{D})$, which contain astronomical knowledge.

Table 3.1. **Class distribution of OGLE labelled set.**

Class	Abbreviation	Number of objects
Long-period variable	LPV	323,999
RR Lyrae	RRLYR	42,751
Eclipsing binary	ECL	41,787
Cepheids	CEP	7,952
Delta Scuti	DSCT	2,807
Type II Cepheids	T2CEP	589
Double periodic variable	DPV	135
Anomalous Cepheid	ACEP	81
Dwarf nova	DN	35
R CrB variable	RCB	22

The desired match between the samples from the proposal and those from the posterior is managed through a function, which is named the bridge function,

$$h(\theta) = C\frac{1}{s_1 p(\mathcal{D}|\theta)p(\theta) + s_2 p(\mathcal{D})g(\theta)}, \tag{3.2}$$

which plays a central role in the bridge sampling estimator (Meng and Wong, 1996). When the bridge function is introduced to the estimator, the function depends recursively on $p(\mathcal{D})$; hence, for estimating it, it is solved recursively by

$$\hat{p}(\mathcal{D})^{t+1} = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2}\frac{p(\mathcal{D}|\theta_i)p(\theta_i)}{s_1 p(\mathcal{D}|\theta_i)p(\theta_i)+s_2\hat{p}(\mathcal{D})^t g_(\theta_i)}}{\frac{1}{N_1}\sum_{j=1}^{N_1}\frac{g(\theta_j)}{s_1 p(\mathcal{D}|\theta_j)p(\theta_j)+s_2\hat{p}(\mathcal{D})^t g_(\theta_j)}}, \tag{3.3}$$

$$\theta_j \sim p(\theta|\mathcal{D}); \theta_i \sim g(\theta). \tag{3.4}$$

Through this estimator, astronomical knowledge is incorporated into the assessment process. Specifically, priors influence the posterior distribution so that it is able to reduce the effect of biases in the training sets.

3.3. Data and classifiers

This section presents the inputs for validating our methodology. Section 3.3.1 describes OGLE catalogue. Section 3.3.2 describes how we obtain the final training set from the crude light curves. Section 3.3.3 explains the procedure to obtain a ground truth.

Lastly, in section 3.3.4, we present the set of models which are assessed through the experiments.

3.3.1. OGLE-III catalogue of variable stars

We use the OGLE-III variable star catalogue for experimental purposes, which corresponds to the third phase of the project (Udalski et al., 2008). The main goal of OGLE project was to identify microlensing events and transiting planets in four fields: the Galactic Bulge, the Large and Small Magellanic Clouds, and the constellation of Carina. We use light curves with at least 25 observations in the I band. The final number of labelled light curves is 420,126. Table 3.1 presents the number of objects per class.

We also use the OGLE-III catalogue to extract the informative priors (step 1 in Figure 3.1). When applying the DRs to this dataset, we obtained $\sim$75% of RR Lyrae stars, $\sim$20% of Eclipsing binaries and $\sim$5% for the rest of the classes.

3.3.2. Processing of light curves

To extract features from the light curves, we use the the feature analysis for time series (FATS) library (Nun et al., 2015), thus obtaining a matrix of 420,126x63, where 63 stands for the number of separate features included in our analysis. Subsequently, in order to manage both dimensionality and multicollinearity, we apply principal component analysis (PCA). Spyroglou et al. (2018) applied a similar strategy that combines Bayesian logistic regression (BLR) and PCA to avoid multicollinearity among features.

To generate a set of models $\mathcal{M}$, we apply linear transformations using PCA with the following set of dimensions: $2, 4, 6, 8, 10, 12$. After this, we apply two polynomial transformations, specifically degrees $1, 2$. This processing allows us to control the complexity of the model either by increasing the number of PCA components or by raising the degree of the polynomial transformation.

3.3.3. Shifted training and testing sets

To evaluate the performance of our approach, we simulate a realistic scenario where the training objects are shifted from the testing objects. In order to create this scenario, we propose splitting a labelled catalogue (OGLE-III in our case) into two shifted (biased) datasets. The main idea to introduce biases is related to separate by the hardness of classification; for that, firstly, we fit a binary classifier (m) that is trained with the entire catalogue. Subsequently, we select objects that are easily separable ($\mathcal{D}_{\text{train}}$) from more difficult objects ($\mathcal{D}_{\text{test}}$).

The idea before commented is exposed in Algorithm 3.3.3. In this algorithm, first of all, we use a random forest classifier (m) to obtain a soft prediction (probability), and then, we use these predictions for splitting the dataset $\mathcal{D}$ as follow. We define a threshold to assess whether an object can be easily classified or not; to measure that, we use the following equation for each object $i \in \mathcal{D}$, $h_i = 1 - (P(A)_i^2 + P(B)_i^2)$, which is based on the Gini impurity index (Raileanu and Stoffel, 2004), being $P(A)_i^2$ the soft prediction for the true-class and $P(B)_i^2 = 1 - P(A)_i^2$, the prediction for the false-class.

To avoid a hard threshold when deciding the set (training or testing) for each object, we add a random selection, which is tuned by a constant T. It is based on the annealing principle (Van Laarhoven and Aarts, 1987) and allows us to provide a probabilistic selection of objects, assigning hard objects ($P(A)_i$ close to 0.5) more frequently to the testing set.

Algorithm 3.3.2 Procedure to bias a catalogue

Input: Data $\mathcal{D} = (\mathcal{D}_{\mathcal{X}}, \mathcal{D}_{\mathcal{Y}})$, classifier m, bias control (T)
Output: Biased Data $\hat{\mathcal{D}} = (\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}})$

```
m.fit(D_X, D_Y)
for (d_x, d_y) ∈ (D_X, D_Y) do
    P_A, P_B ← m.softPredict(d_x)
    h = 1 − (P_A^2 + P_B^2)
    p = e^{−h/T}
    r = uniform(0, 1).sample()
    if r ≤ p then
        D_train.add((d_x, d_y))
    else
        D_test.add((d_x, d_y))
    end if
end for
return D̂
```

We apply Algorithm 3.3.3 in order to obtain datasets with different levels of bias for the RR Lyrae class. The bias was managed by the parameter T, and we obtained datasets *rrlyrae-1*, *rrlyrae-2* and *rrlyrae-3* for $T \in \{1, 2, 4\}$ as a result. These three configurations allow us to evaluate our proposal under different bias scenarios. Table 3.2 provides a summary of different biased datasets.

Figure 3.2 shows the distribution of the hardness (classification difficulty) for objects in training and testing sets. As we said before, we assume that predictions close to 0.5 are more difficult to classify than predictions close to 1 or 0. According to this definition, in the training sets in Figures 3.2(a),3.2(c) and 3.2(e), we can observe that the training sets have higher frequency of easier objects than the testing sets in Figures 3.2(b),3.2(d) and 3.2(f). The relative frequency of objects in different levels of hardness can be visualised in both the histograms and those bars that represent objects on the top of each figure.

From a further perspective, we can see Figure 3.3 that presents the resulting biases, by using Algorithm 3.3.3, in the space of features for dataset *rrlyrae-1*. In particular, in Figures 3.3(a) and 3.3(b), we present objects located at the Small Magellanic Cloud; it is clear a data shift in the joint density distribution (amplitude and period) between the

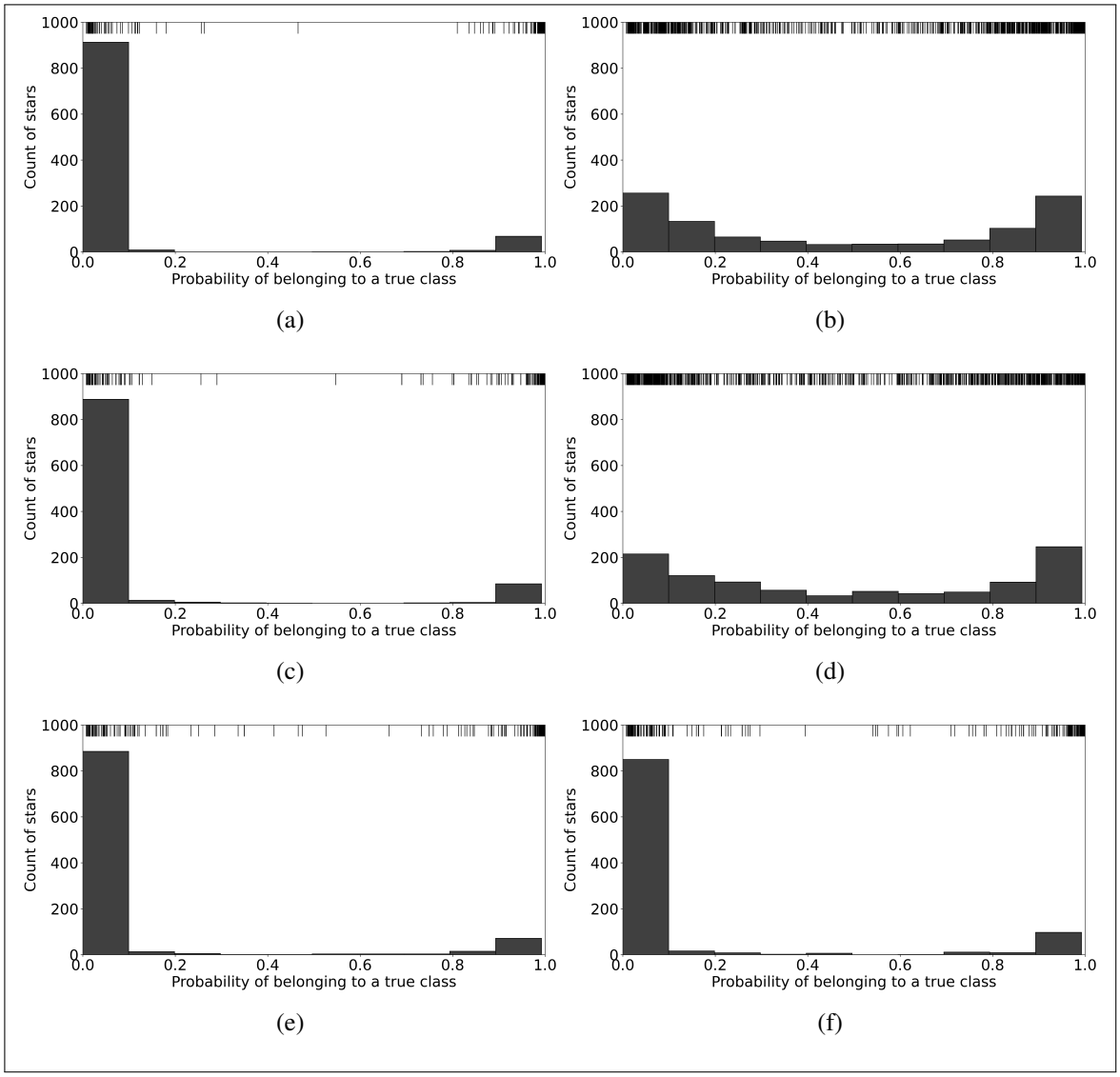


Figure 3.2. **Histogram for the probability of belonging to the true class in biased datasets.** Figures (a) and (b) illustrate *rrlyrae-1*; Figures (c) and (d) show *rrlyrae-2*; and Figures (e) and (f) present *rrlyrae-3*. Figures (a), (c), and (e) correspond to the training sets, while Figures (b), (d), and (f) correspond to the testing sets. The bars at the top of each figure indicate objects. A sample of 1,000 objects was used to generate these plots in each set.

training and testing sets. Moreover, Figures 3.3(c) and 3.3(d) show that in the Galactic Disk Field also was generated a shift between training and testing sets.

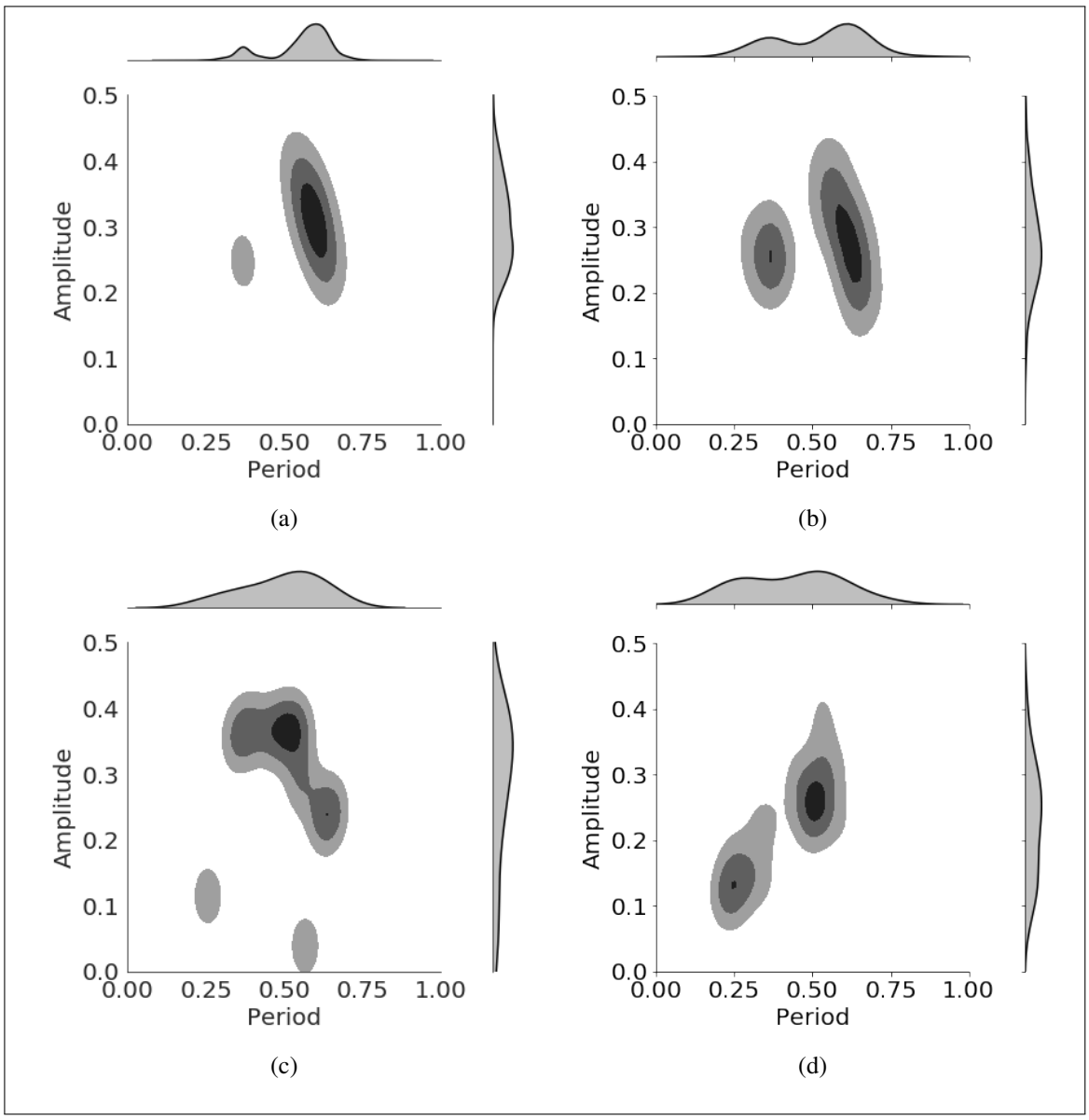


Figure 3.3. **Density plots for RR Lyrae variable stars in *rrlyrae-1* dataset.** (a) Small Magellanic Cloud - Training. (b) Small Magellanic Cloud - Testing. (c) Galactic disk - Training. (d) Galactic disk - Testing.

Table 3.2. **Number of objects in training and testing for each class.** TC represents true-class.

$\mathcal{D}$	Training	Testing	TC training	TC testing
rrlyrae-1	389,364	30,762	27,240 (6.9%)	15,269 (49.6%)
rrlyrae-2	402,787	17,339	34,233 (8.5%)	8,500 (49.0%)
rrlyrae-3	335,721	84,405	34,001(10.1%)	8,732(10.3%)

3.3.4. Classifiers

As we mentioned before, we focus on Bayesian logistic regressions, although the aforementioned method can be applied to other Bayesian classifiers. Furthermore, we compare the informative marginal likelihood ranking with the ACC-based ranking in a CV framework for two traditional logistic regression variants. Below, we present a brief description of each of these.

Standard logistic regression (LR): The standard LR classifier models the success probability of a binary dependent variable, $y \in \{0, 1\}$ by means of a Bernoulli distribution,

$$p(y|x, \theta) = \text{Ber}(y|s(x, \theta)). \tag{3.5}$$

In this model, a sigmoid function,

$$s(x, \theta) = \frac{1}{1 + e^{-\theta^T x}} = \frac{e^{\theta^T x}}{1 + e^{\theta^T x}}, \tag{3.6}$$

of inputs ($\mathbf{X}$) and parameters (θ) is used for modelling the bernoulli parameter ($p = s(x, \theta)$). Finally, the likelihood function,

$$p(y|x, \theta) = s(x, \theta)^y (1 - s(x, \theta))^{1-y}, \tag{3.7}$$

is optimised, giving rise to the maximum likelihood estimator.

Penalised logistic regression (l_2-LR): In Bayesian terms, penalised LRs (l_2 and l_1) embody a prior distribution over θ, and subsequently, the maximum value for the resulting distribution (maximum a posteriori or MAP) is selected. In particular, l_2-LR is equivalent to a vague Gaussian prior centred at the origin. This approach does not use human knowledge to define the shape of priors.

Bayesian logistic regression (BLR): BLR focuses on estimating and using the posterior of distribution for the weights $p(\theta|\mathcal{D})$ in the LR. In our proposal, the informative priors $\sim \mathcal{N}(\hat{\theta}, \hat{\sigma})$ are estimated using the method laid out in Section 3.2. For these experiments,

we consider DRs for period and amplitude; both were estimated with FATS library Nun et al. (2015).

Each of these models (LR, l_2-LR and BLR) represents a family of models ($\mathcal{M}$), which are defined by transformations over their input matrix. Regarding these transformations, as was commented before, firstly, we apply a dimensional reduction using principal component analysis, and after that, we apply a polynomial transformation. Let ($\mathbf{X}^{\mathrm{n}\times\mathrm{m}}$) be the final matrix, being n the objects in the training set and m the product between the polynomial degree and the number of components used in each model $m \in \mathcal{M}$. Then, our goal is to assess and sort the models within each family. Downsampled data were used to deal with imbalanced classes. The notation model(a, b) indicates a model with a polynomial transformation of degree a and dimensionality reduction to b components.

In LR and l_2-LR, the models are sorted by their cross-validated ACC in training. The BLR models are ordered according to our method (informative marginal likelihood). In the following section, we compare the rankings provided and show empirical results in which the marginal likelihood can provide improved rankings compared to the CV (LR and l_2-LR).

3.4. Results

This section compares our informative BMS strategy with standard methods based on non-informative CV. Figure 3.4 presents two examples of rankings that were generated by different strategies for selecting models: Figure 3.4(a) provides a ranking of models from our proposed method, while Figure 3.4(b) shows a ranking using a k-fold cross-validated ($k = 10$) ACC. In these simple examples, we can note that the marginal likelihood provides a better ranking correlation compared to the cross-validated ACC. The significance of this is that the marginal likelihood is more robust to bias in the training set.

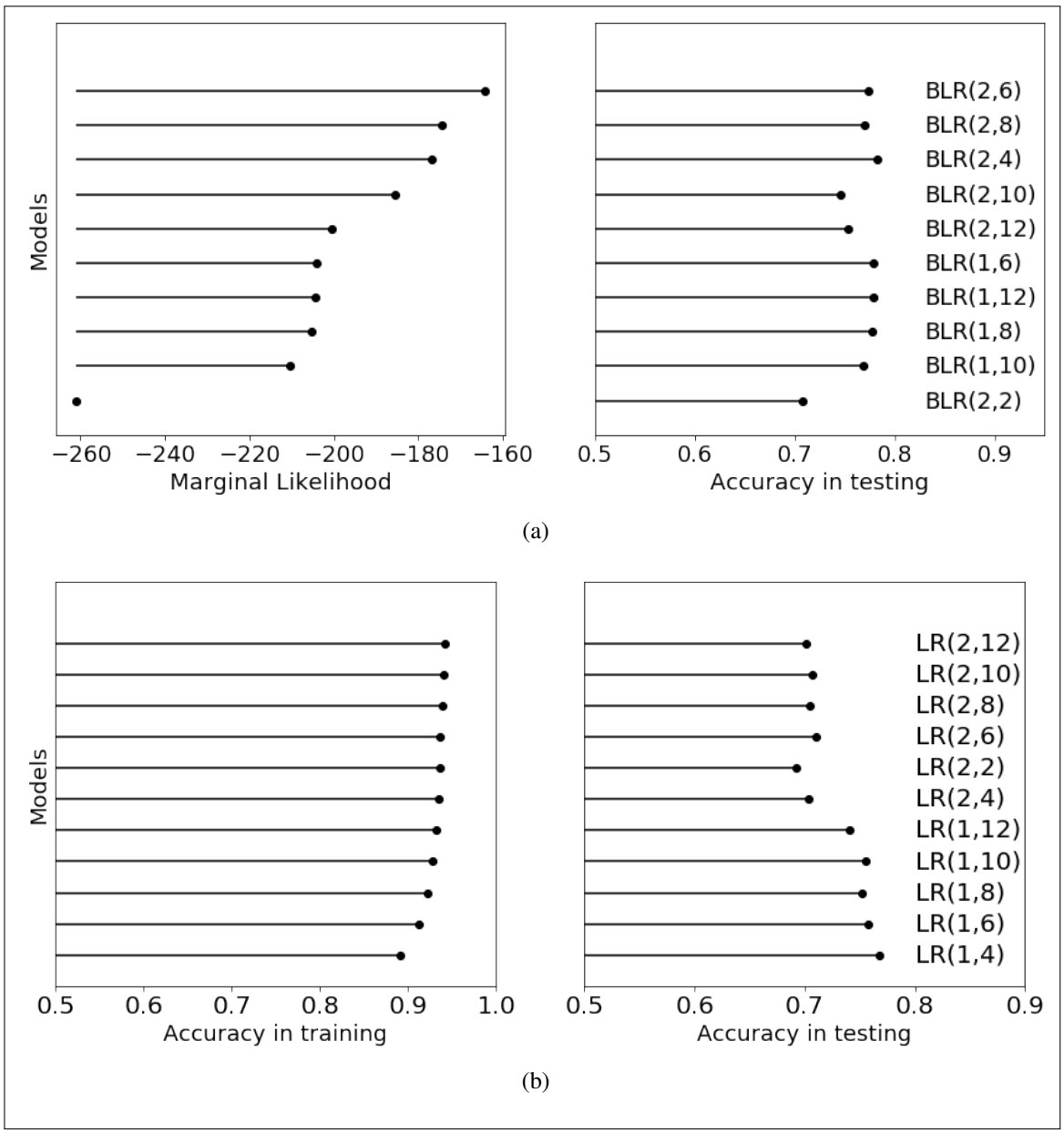


Figure 3.4. **Comparison of model rankings with 1,000 samples on *rrlyrae-3* set.** (a) Models are sorted by the marginal likelihood **BLR-IP**(σ=10). (b) Models sorted by a cross-validated (k=10) ACC for l_2-**LR**-1 family of models.

To obtain a more rigorous comparison of rankings among methods, we define a set of metrics to quantify several viewpoints thereof. The selection of metrics used to compare models is presented below.

Kendall-tau (τ) correlation this metric is estimated by, $\frac{n_c - n_d}{\frac{1}{2}n(n-1)}$, where n_c represents the number of concordant models and n_d is the number of discordant models in the ranking. Identifies the coincidences between training and testing rankings. Both rankings (training and testing) are correlated when the selection is correct. In preliminary experiments, similar results were obtained by using the Spearman's rank correlation coefficient.

ACC score: To discriminate beyond the ranking order, we measure the accuracy rate of the foremost three models for each ranking. Thus, we are able to identify the quality of the selected models.

F_1 **score:** To identify the impact of unbalanced classes, we estimate the average F_1-score in testing for the three foremost models in the ranking.

Delta training/testing (Δ_T): Seeks to evaluate how far the predicted ACC rate for the testing set is with respect to the training set.

In tables 3.3-3.5, we summarise the metrics and datasets (*rrlyrae-1*, *rrlyrae-2* and *rrlyrae-3*) for three subset training sizes s. We run three baseline strategies to rank models in addition to three approaches based on the marginal likelihood. The standard approaches (LR, l_2-LR-1 and l_2-LR-100) are based on k-fold CV using $k = 10$. The rankings based on the marginal likelihood consider flat priors (BLR-FP); information on mean and fixed variance (BLR-IP $\sigma = 10$); and information on mean and variance (BLR-IP).

We can observe that the τ correlations are greater for marginal likelihood-based rankings compared to CV-based rankings. This demonstrates that the marginal likelihood is more robust than the cross-validated ACC for addressing different levels of bias. Specifically, with regard to *rrlyrae-1* and *rrlyrae-2* data the best rankings were provided by BLR-IP (2,000), while the best ranking for *rrlyrae-3* was obtained using BLR-IP (1,000).

Concerning F_1-score and ACC, the three best informative Bayesian models obtain a better performance than the three best likelihood-based models. This means that the predictive performance of posterior samples (posterior mean) is also improved when prior

Table 3.3. **Evaluations of rankings of models in dataset *rrlyrae-1*.** τ is the Kendall's tau rank correlation; ACC and F1 are the mean accuracy rate and the mean F1-score, respectively; of the three foremost models, Δ_T is the average difference between the ACC in training and testing. The bold numbers represent the best strategy for model selection by each metric.

k-fold CV					
$\mathcal{M}$	s	τ	F_1	ACC	Δ_T
LR	1,000	0.11	0.62	0.62	0.35
	2,000	0.09	0.62	0.63	0.34
	4,000	0.26	0.63	0.65	0.32
l_2-LR-100	1,000	-0.31	0.64	0.64	0.35
	2,000	-0.09	0.64	0.63	0.35
	4,000	0.24	0.68	0.66	0.32
l_2-LR-1	1,000	-0.16	0.64	0.64	0.35
	2,000	-0.02	0.63	0.63	0.34
	4,000	0.49	0.69	0.66	0.32
marginal likelihood					
$\mathcal{M}$	s	τ	F_1	ACC	Δ_T
BLR-FP	1,000	0.70	**0.70**	**0.69**	0.32
	2,000	0.82	**0.70**	**0.69**	0.32
	4,000	**0.85**	**0.70**	**0.69**	0.31
BLR-IP (σ=10)	1,000	0.31	0.68	**0.69**	0.32
	2,000	0.56	**0.70**	**0.69**	0.32
	4,000	0.75	**0.70**	**0.69**	0.31
BLR-IP	1,000	0.60	0.50	0.61	**0.24**
	2,000	**0.85**	0.65	0.66	0.34
	4,000	0.71	0.69	0.68	0.32

knowledge from DRs is added. When looking at experiments with *rrlyrae-1*, it is worth noting that the difference is more significant at the ACC metric. When comparing the Δ_T, a smaller difference between training/testing performances is observed in *rrlyrae-1*. Significant differences are not obtained in *rrlyrae-2* and *rrlyrae-3*.

Finally, Table 3.5 demonstrates that, although high levels of ACC are achieved, we observe small (or even negative) correlations in *rrlyrae-3*, along with a low F1-score. Notably, the BLR-IP model uses 1,000 objects, which enables it to perform best across all five metrics. Additionally, we emphasise that this highly informative setting leads to better performance during testing compared to training, with a performance difference of Δ_T = -0.25.

Table 3.4. **Evaluations of Rankings of Models in dataset *rrlyrae-2*.** Columns represent the same information as in Table 3.3.

k-fold CV					
$\mathcal{M}$	s	τ	F_1	ACC	Δ_T
LR	1,000	0.13	0.67	0.64	0.32
	2,000	0.38	0.66	0.66	0.32
	4,000	0.38	0.63	0.65	0.32
l_2-LR-100	1,000	0.27	0.65	0.64	0.32
	2,000	0.24	0.65	0.63	0.33
	4,000	0.45	0.65	0.63	0.34
l_2-LR-1	1,000	0.42	0.66	0.65	0.32
	2,000	0.27	0.65	0.63	0.33
	4,000	0.45	0.65	0.63	0.34
marginal likelihood					
$\mathcal{M}$	s	τ	F_1	ACC	Δ_T
BLR-FP	1,000	0.71	0.68	0.66	0.32
	2,000	0.64	0.68	0.66	0.33
	4,000	0.76	**0.69**	**0.67**	0.32
BLR-IP (σ=10)	1,000	0.20	0.68	**0.67**	0.33
	2,000	0.61	0.68	0.66	0.33
	4,000	0.73	0.68	**0.67**	0.32
BLR-IP	1,000	0.60	0.49	0.59	**0.24**
	2,000	**0.82**	0.68	0.66	0.33
	4,000	0.78	0.68	**0.67**	0.33

Table 3.5. **Evaluations of Rankings of Models in dataset *rrlyrae-3*.** Columns represent the same information as in Table 3.3.

k-fold CV					
$\mathcal{M}$	s	τ	F_1	ACC	Δ_T
LR	1,000	-0.75	0.46	0.76	0.19
	2,000	-0.53	0.45	0.75	0.20
	4,000	-0.05	0.57	0.85	0.12
l_2-LR-100	1,000	-0.75	0.46	0.76	0.20
	2,000	-0.78	0.44	0.74	0.22
	4,000	-0.42	0.51	0.80	0.16
l_2-LR-1	1,000	-0.71	0.46	0.76	0.20
	2,000	-0.78	0.44	0.74	0.22
	4,000	-0.49	0.51	0.80	0.17
marginal likelihood					
$\mathcal{M}$	s	τ	F_1	ACC	Δ_T
BLR-FP	1,000	-0.16	0.49	0.78	0.19
	2,000	-0.36	0.46	0.76	0.22
	4,000	-0.39	0.46	0.76	0.21
BLR-IP (σ=10)	1,000	0.09	0.49	0.79	0.18
	2,000	-0.30	0.47	0.76	0.22
	4,000	-0.33	0.46	0.76	0.21
BLR-IP	1,000	**0.56**	**0.71**	**0.93**	**-0.25**
	2,000	-0.53	0.47	0.76	0.22
	4,000	-0.53	0.46	0.76	0.22

4. CHAPTER IV. INFORMATIVE REGULARISATION FOR A MULTI-LAYER PERCEPTRON RR LYRAE CLASSIFIER UNDER DATA SHIFT

4.1. Method overview

This section introduces our second contribution: a methodology to mitigate the data shift problem by regularizing an MLP model using expert knowledge. First, we collect deterministic rules of ranges for characteristic features to construct a symbolic representation of prior knowledge, which was used to model the informative regulariser component. Simultaneously, we implement an ad-hoc training procedure. We design a two-step back-propagation algorithm to integrate this knowledge into the neural network, whereby one step is applied in each epoch to minimise classification loss, while another is applied to ensure regularisation. Our algorithm defines a mask, which is a subset of parameters, for each loss function. This approach mitigates the forgetting effect, which stems from a trade-off between these functions (overfitting to data versus learning expert knowledge) during training.

4.2. Knowledge injection

In the scenario introduced by this proposal, training data are shifted and, thus, standard MLP learns from those shifted data. If traditional regularisation techniques are applied without human knowledge, the problem remains. Since learning takes place on a non-representative set, only relevant patterns of shifted data are identified. Our approach proposes using an informative regularisation approach based on the symbolic representation of well-documented physical knowledge of variable stars to mitigate this data shift problem. This representation is based on characteristic intervals of values for an invariant feature i of a variable star a, henceforth denoted as r_i^a, that can be defined by:

$$\text{class } a \rightarrow \ x_i \in [l_i, u_i],$$

where l_i and u_i are the lower and upper characteristic limits for feature i. Below we present a set of examples (e.g., Catelan and Smith, 2015, and references therein):

- RR Lyrae $\Rightarrow$ (period $\in [0.2, 1.0]$ days)
- RR Lyrae $\Rightarrow$ (amplitude $\in [0.3, 1.2]$ in V-band)
- RR Lyrae $\Rightarrow$ (amplitude $\in [0.2, 0.8]$ I-band)
- Classical Cepheid $\Rightarrow$ (period $\in [1, 100]$ days)
- Type II Cepheid $\Rightarrow$ (period $\in [1, 30]$ days)

To inject expert knowledge into the MLP training, we propose to use synthetic inputs called signals. A signal is defined as a sparse vector (synthetic training data) with non-zero components associated with a set of DRs. Signals contain information about characteristic and invariant features; the objective herein is to draw samples $s(i)$ for each feature i from a proposed distribution q, the parameters of which depend on r_i^a. For example, based on the assumption that q has a uniform distribution, and if we include a perturbation ϵ on the limits, the following samples can be generated:

- $\underline{s}(i) \sim \mathcal{U}(l_i - \epsilon, l_i + \epsilon)$
- $\bar{s}(i) \sim \mathcal{U}(u_i - \epsilon, u_i + \epsilon)$

Each signal is represented by a sparse input vector $\mathbf{s} = [0, 0, .., s(i), .., 0, 0]$. The data set that contains those signals is denoted $\mathcal{D}^{\mathbf{s}}$. $s(i)$ contains knowledge only from the DR borders since, based on preliminary experiments, populating the entire DR range does not improve classification performance and requires a considerably larger computational effort. Therefore, we focus on the DR limits, which are the less represented zones in unidimensional distributions.

To generate a bidimensional (2D) signal based on two DRs, class $a \rightarrow (x_i \in [l_i, u_i]) \wedge (x_j \in [l_j, u_j])$, we propose the following two alternatives for q: (i) independent unidimensional (1D) samples can be generated and (ii) a fitting over a bivariate Gaussian using a

subset of data based on DRs, i.e., observation close to the DR borders. In the latter alternative, to mitigate the data shift problem, we apply a shift $\Delta\mu$ over the mean, and we assume a full-matrix covariance.

In this case, the signal is $\mathbf{s} = [0, 0, .., s(i), .., s(j).., 0, 0]$. To select the subset of data, we propose using objects from the class (RR Lyrae); and subsequently, we select a subset of objects from the intersection of the DR bounds. In the case of two rules, the four valid filtering options are presented below.

a) $(x_i \leq l_i^a + \epsilon_s) \wedge (x_j \leq l_j^a + \epsilon_s)$
b) $(x_i \geq u_i^a - \epsilon_s) \wedge (x_j \geq u_j^a - \epsilon_s)$
c) $(x_i \leq l_i^a + \epsilon_s) \wedge (x_j \geq u_j^a - \epsilon_s)$
d) $(x_i \geq u_i^a - \epsilon_s) \wedge (x_j \leq l_j^a + \epsilon_s)$

The ϵ_s hyperparameter manages the distance from the filtered data to the characteristic DR. Our experiments consider the use of filters a) and b), and the ϵ_s parameter was set in the range [0.2, 0.4]. Note that, when few objects (i.e., ~ 100) are obtained to fit the Gaussian, ϵ_s must be increased; we recommend an increment of 0.02 each time.

Using sparse signals helps inject expert knowledge of characteristic features in underrepresented data zones. Moreover, when signals contain few dimensions, human knowledge can be more easily incorporated into proposal distributions.

Traditional approaches for controlling biases or class imbalance in data based on data augmentation can need to generate large samples of new objects to be effective. In our approach, the training of a subset of weights (a mask) is proposed, whose responsibility is to manage and transmit human knowledge through the neural network. The sparse signals do not need to compete for resources (ANN learning capacity) with training data since the MLP contains a set of weights focused on learning this knowledge. The following section will explain this ad-hoc modelling and the training procedure.

4.3. Training procedure

This section describes an ad-hoc procedure that optimizes classification loss and sets structural knowledge from the DRs. Our proposed training procedure is composed of the following elements:

Bi-objective MLP: We optimize an MLP considering it as a standard bi-objective modelling as follows:

$$\bar{\mathcal{L}} = (\mathcal{L}_1, \mathcal{L}_2), \tag{4.1}$$

$$\min \mathcal{L}_1, \tag{4.2}$$

$$\min \mathcal{L}_2, \tag{4.3}$$

where $\mathcal{L}_1$ is the binary cross-entropy loss function for training cases, and $\mathcal{L}_2$ is a binary cross-entropy for signals. In non-shifted data, there is no clear trade-off between the information contained in the data and knowledge from experts; however, in the presence of a data shift problem, we show that structural rules can induce a trade-off and mitigate biases.

This approach is different to traditional approaches, whose loss functions are weighted as follows:

$$\min z = \alpha\mathcal{L}_1 + (1-\alpha)\mathcal{L}_2, \tag{4.4}$$

where $0 \leq \alpha \leq 1$. In fact, in the weighting sum approach, for each α-value in equation (4.4) we have a solution in the Pareto frontier (non-dominated solutions) for the global bi-objective problem in equations (4.1)-(4.3) (Ehrgott, 2005). In equation (4.4), the balance between objective functions is managed by α, and sometimes it can be difficult to find the zone where the trade-off in the objectives is generated. In our case, the balance is managed by a hyperparameter, ϵ_w, which is a simple threshold of values of weights for each loss; hence, this is a more intuitive and controllable approach for assigning priority to each objective depending on the bias level.

Consequently, a learning procedure is designed based on two objective functions. The first, $\mathcal{L}_1$, considers a standard classification loss (e.g., cross-entropy). The second, $\mathcal{L}_2$, is focused on fitting the high-level knowledge to obtain a more robust boundary. For $\mathcal{L}_2$, preliminary experiments are conducted using negative, positive and ambiguous signals (probability equal to 0.5). The most significant improvements were obtained by applying a mixture of positive signals (6,000) and negative samples (6,000); therefore, these results are reported.

Two-step back-propagation: A two-step back-propagation algorithm is applied to set the human knowledge in the neural network, in which one step is used for a traditional loss (e.g., cross-entropy) and the other for the regularisation loss. Batches for regularisation loss contain informative signals and negative examples (objects other than RR Lyrae) from the training set. Adam optimiser was applied for both losses, in which the learning rate for the primary loss was twice that of the auxiliary loss. To avoid the gradient exploding problem, we apply a norm gradient clipping (Zhang et al., 2019).

Masks: An iterative back-propagation is implemented with masks focused on specific losses, i.e., iterations 1, 3, 5, . . . , $2k+1$ are focused on optimising $\mathcal{L}_1$ and iterations 2, 4, 6, . . . , $2k$ are focused on $\mathcal{L}_2$. Each loss has a binary mask for the back-propagation step. The masks can be understood as two independent systems from a modelling perspective. A primary system, which is based on training data, aims to memorise the class for each data belonging to the training set. A secondary system is responsible for incorporating structural knowledge into the learning process. An initial training phase is performed to assign the weights to each mask (first 250 epochs of training). In this phase, the weights can change of assigned mask (assigned loss function) in each epoch. The weights smaller than a threshold, ϵ_w, are assigned to $\mathcal{L}_2$, and the remaining weights are optimised considering $\mathcal{L}_1$. Smaller weights ($\leq |\epsilon_w|$) are assigned to signal-based loss ($\mathcal{L}_2$); in other words, weights with low relevance for activating or inhibiting neurons are optimised for the auxiliary loss of regularisation. Following this initialisation phase, the weights are fixed to a

loss. Bias parameters (θ_j) and batch normalisation learnable parameters are assigned to the primary system to provide more stable training.

Other hyperparameters: In preliminary experiments, batch sizes of 32, 64, 128, 256 and 512 are used. The best results were obtained with a batch size of 256; hence, only these were reported. ReLU activation functions are always used throughout. An early stopping strategy using a patience parameter equal to ten epochs is implemented, i.e., the validation loss checking is carried out every ten epochs. The maximum number of epochs is 1,500. Due to the imbalance problem in the training set, a downsampling approach is applied. Based on preliminary results, hyperparameter ϵ_w was set to 0.025 and hyperparameter ϵ_s was set to 0.2.

Figure 4.1 provides a visualisation of the training scheme, in which the features that receive human-based information from signals are coloured red. Green circles represent activation functions. Connections (weights) between nodes that are optimised via the application of the back-propagation from $\mathcal{L}_2$ are coloured red. The remaining weights are updated by using back-propagation from $\mathcal{L}_1$ and are coloured black. Figure B.1 shows a sample of the convergence behaviour of our training procedure.

4.4. Data

This section presents the fundamental elements for validating the proposal outlined in this chapter. Section 4.4.1 briefly describes the OGLE-III catalogue (Udalski et al., 2008). Section 4.4.2 provides details about how the final data set is obtained from the crude light curves. Finally, Section 4.4.3 explores a shifted data set, which was subsequently used to experiment and validate the applied methodology.

4.4.1. OGLE-III catalogue of variable stars

The OGLE-III catalogue of variable stars was used to assess the proposed methodology. As its name implies, the catalogue corresponds to the project's third phase (Udalski et

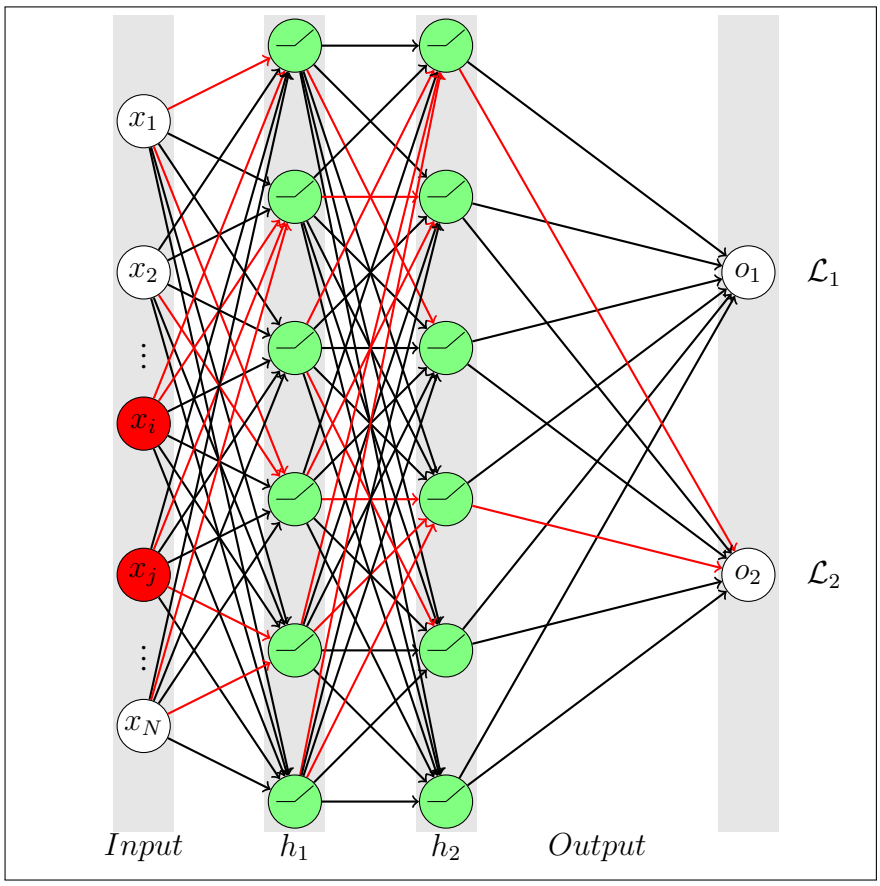


Figure 4.1. **Bi-objective MLP with masks for each objective.** Each grey rectangle contains a layer. In the input layer, red circles represent the non-zero values in each signal. Green circles in hidden layers h_1 and h_2 represent neurons. Arrows show the flow of information through the network. The arrow colour indicates the loss function from where the weight update comes, where $\mathcal{L}_1$ represents the primary objective, which is related to the classification task, and $\mathcal{L}_2$ is the regularisation loss, which is focused on adding human knowledge into the training. Red weights are updated using the propagated error from $\mathcal{L}_2$, and black weights are updated using the back-propagation from $\mathcal{L}_1$.

al., 2008). The main goal of OGLE is to identify microlensing events and transiting exoplanets in four fields: the Galactic bulge, the Large and Small Magellanic Clouds, and the constellation of Carina. This study used OGLE light curves with at least 25 observations in the I band.

4.4.2. Processing of light curves

The corresponding features of each light curve were extracted using the FATS library (Nun et al., 2015). A sample of 300 observations was used for each light curve with more

than 300 observations. Objects with outliers (i.e., values beyond the mean $\pm$ three standard deviation) in features were discarded. Thus, a matrix of 329,684$\times$60 was obtained, where 60 stands for the number of separate features included in the analysis. These features will be used to experiment with the aforementioned training procedure for MLPs.

4.4.3. Shifted data

To validate the approach used in this study for mitigating the data shift problem, we apply the method for shifting data proposed in Section 3.3.3. This benchmark set was designed to emulate the underlying biases in variable star training data. Following a description of the aforementioned pre-processing, this section moves on to discuss the distribution of variable star types, which is presented in Table 4.1.

Two levels of information are visualised for both sets. On the first level, Figure 4.2 shows a sample of light curves belonging to the training and testing sets. On the second level, Figure 4.4 contrasts training and testing sets over the space of the most relevant features (Nun et al., 2015). The relevance of these features was estimated using decision trees, and the mean decrease was calculated using the impurity method.

When observing Figure 4.2, detecting the shift between light curves belonging to training and testing sets is not straightforward since there are multiple patterns and points in each light curve. Despite this, the data shift is visualised when the feature space is observed in Figures 4.3 and 4.4.

Figure 4.3 provides a visualisation of the period and amplitude for RR Lyrae stars in training and testing sets. In the main-diagonal sub-plots, univariate distributions for period and amplitude are considered. The upper right subplot includes merging objects from both sets and provides the joint density distribution. The subplot on the lower left shows a contour of joint density distribution for each data set, i.e., a density contour for the training set (light blue) and a contour for the testing set (salmon). The shift is clearly identified in

Table 4.1. **Training and testing class distribution from OGLE-III labelled set.**

Class	**Testing**	**Training**
RR Lyrae	12,031	24,169
Eclipsing Binary	5,307	18,517
Cepheids	3,700	2,878
Long-Period Variables	851	260,518
Delta Scuti	718	493
Type II Cepheid	179	158
Anomalous Cepheid	42	31
Double Periodic Variable	11	72
Dwarf Nova	1	0
α^2 Canum Venaticorum	0	4
R CrB Variable	0	4

Table 4.2. **Summary of model's performance decay from the training set to the testing set.** Each experiment represents the mean value on the testing set from a baseline model trained using a random sampling approach within the training set. An early stopping strategy is considered, using a validation set from the training set to avoid overfitting.

	ACC		F1 score		AUC	
	train	test	train	test	train	test
5,000	99.78	77.42	99.87	77.42	1.00	0.84
10,000	99.81	78.87	99.88	79.15	1.00	0.85
50,000	99.87	77.49	99.91	82.98	1.00	0.85

the univariate case; for example, the same two populations are observed in univariate period density, but the balance between modes is shifted. In the case of univariate amplitude density, the mode of the first population and the relative representation of each population are also shifted.

Figure 4.4 provides the same visualisations for the seven most relevant features previously introduced. The shift is identified in each univariate density since the kurtosis is constantly shifted and, in certain cases, e.g. amplitude, the means also differ. In bivariate plots below the main diagonal, most of the density mass is observed in at least two populations.

Consequently, these induced biases generate a performance decay in the trained models. Table 4.2 shows this detriment, considering the following three metrics:

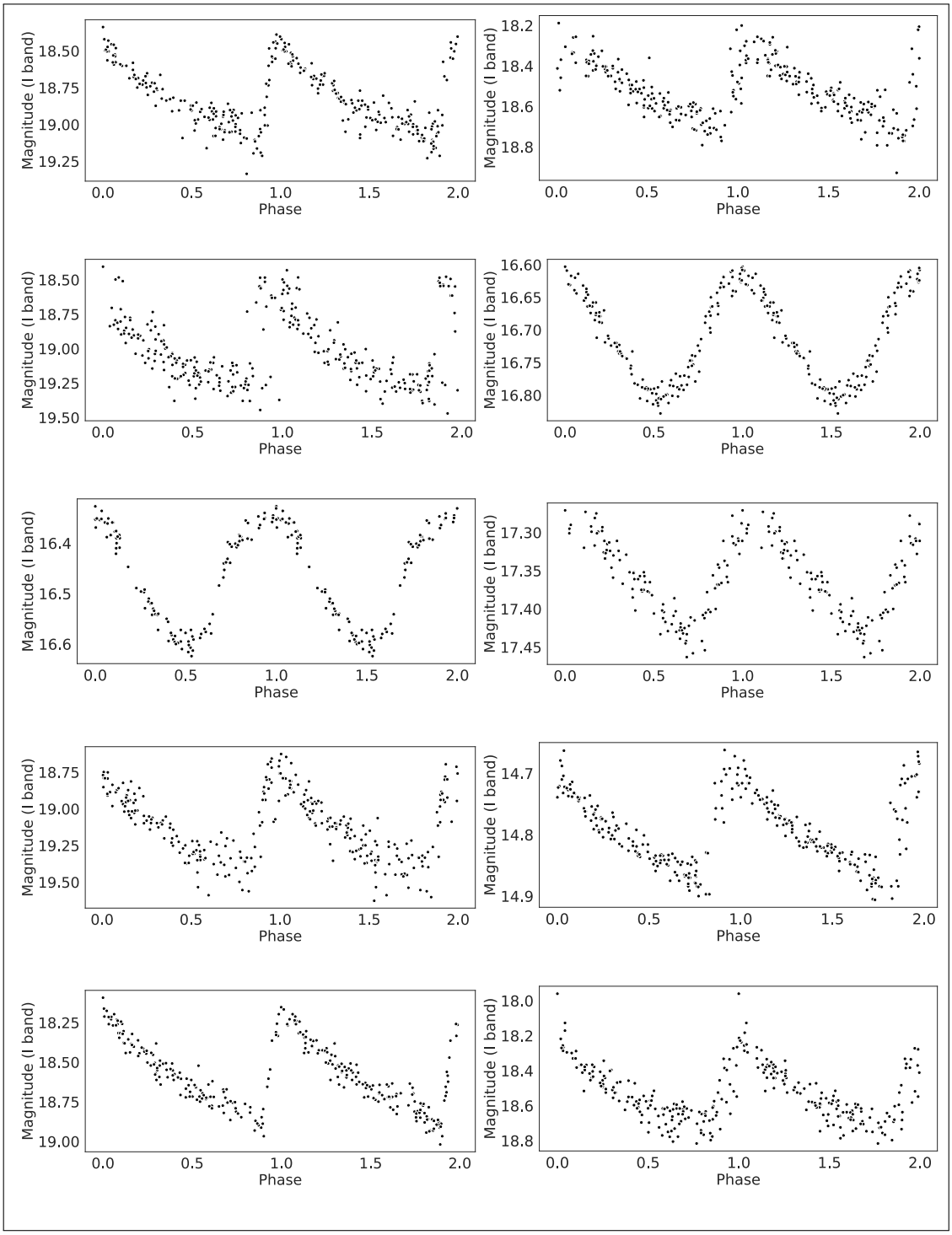


Figure 4.2. **Sample of folded light curves of RR Lyrae variable stars.** The stars to the left relate to the training set, and the stars to the right relate to the testing set.

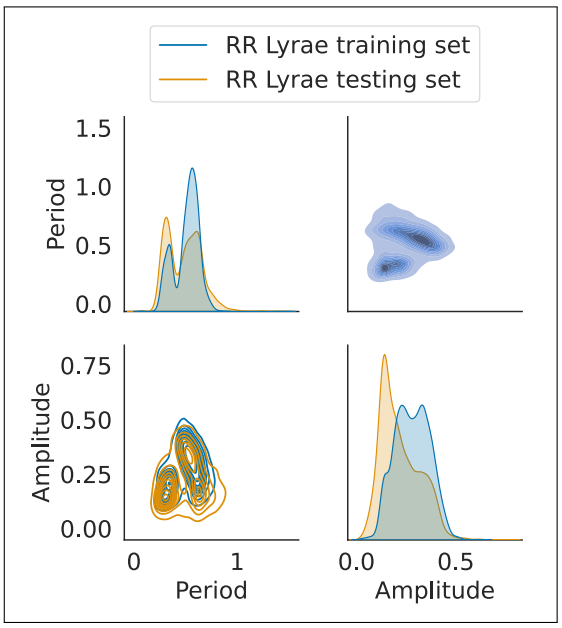


Figure 4.3. **Period and amplitude density distributions for training and testing sets.** The upper left and lower right subplots show the univariate density fit of each feature; the upper right plot shows the bivariate density of these features, merging testing and training sets (coloured blue), whereas the lower left plot shows the bivariate density fit contour for each set. The density contour and shadows are coloured light blue and salmon for the testing and training sets. We apply the kernel density estimation method to provide these visualisations using a sample of 5,000 objects in each data set.

ACC: This evaluates the prediction quality on the testing set based on the ratio of correct predictions over the total number of observations. We consider a threshold equal to 0.5 for defining the true class.

F_1-score: To include a balance between recall (i.e., true positive rate) and precision (i.e., positive predictive rate), we consider the F_1-score, which is the harmonic mean between these; in addition, this is more robust than ACC for imbalanced classes.

AUC: Unlike ACC and F_1 metrics, the area under the receiver-operating characteristic curve (AUC-ROC) is based on the model performance for each possible classification threshold (Davis and Goadrich, 2006). The receiver-operating characteristic curve provides the true positive rate with respect to the false positive rate, being an AUC equal to 1.0 for a perfect model score and 0.5 in the case of a random model. AUC can be understood as the probability that the assessed model classifies a random true-class object more probably than a random false-class object.

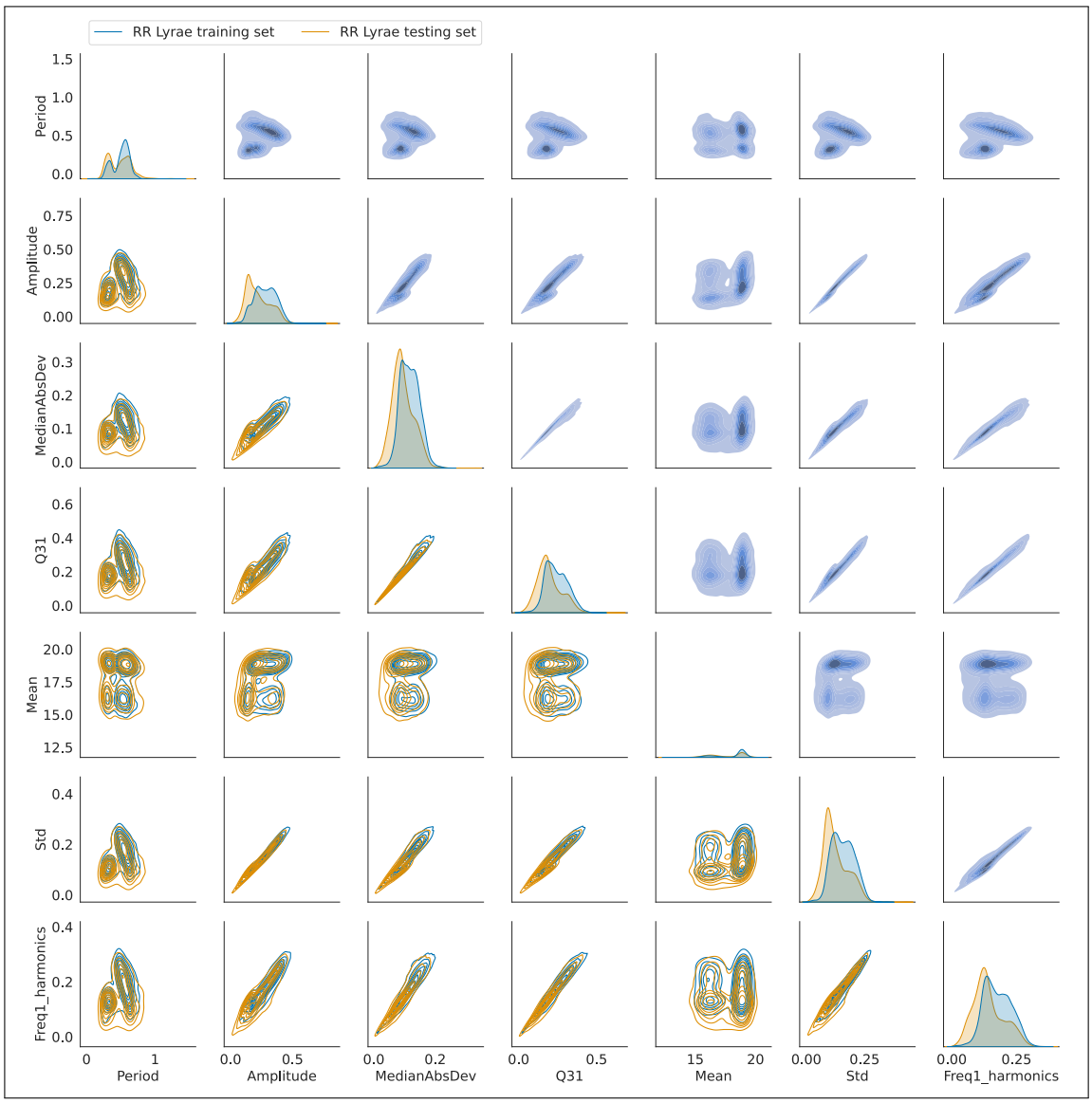


Figure 4.4. **Probability density distribution for a subset of features.** The main-diagonal subplots show a univariate density fit of each feature; the plots above the main-diagonal show the bivariate density of these features, merging testing and training sets, whereas the plots below the main-diagonal show the bivariate density fit contour for each set.

As we can see, the model's performance in the three metrics is perfect on training sets, getting the maximum score (or very close to the maximum). However, these metrics decay when the model is tested beyond the training set, even when traditional strategies are applied to mitigate overfitting. The baseline model used in these experiments is an

MLP with the hyperparameters (and architecture) presented in Section 4.3 but without our informative regularisation approach.

4.5. Experiments and results

This section presents the experimental results. Section 4.5.1 discusses the use of 1D signals, assessing the impact of our approach in terms of classification performance on the shifted RR Lyrae data presented earlier (Section 4.4). Section 4.5.2 shows the results of injecting 2D signals and also provides a comparison with respect to traditional non-informative regularisation approaches.

4.5.1. Regularisation using unidimensional knowledge injection

To assess the impact of our proposal, when 1D signals are injected during the MLP training, we compare the following four approaches: (i) an MLP without regularisation, the baseline model; (ii) the baseline model using dropout as regularization; and two (iii-iv) informative regularisation approaches using our training procedure. Our proposal includes signals based on expert knowledge from period and amplitude features, which we refer to as 1D-period and 1D-amplitude.

Figure 4.5 shows the distribution of the test-set ACC for 30 experiments. When the informative loss is activated, we note a bigger concentration towards high ACC values. Moreover, the mean ACC is also improved (from 77.9% to 78.9%) when the informative loss is used. We apply the t-student[1] tests for independent samples to validate these experiments statistically. In both cases, the null hypothesis was rejected ($\alpha = 0.05$), thus implying that the differences in test-set ACC are statistically significant.

Figure 4.6 shows the impact of our regularization method injecting 1D signals on the ACC. White points represent the mean, black lines show the median, and dotted lines

[1] A Shapiro-Wilk test was used to validate normality. We also apply a Mann-Whitney-Wilcoxon test to check statistical significance, showing similar results with respect to the t-student test.

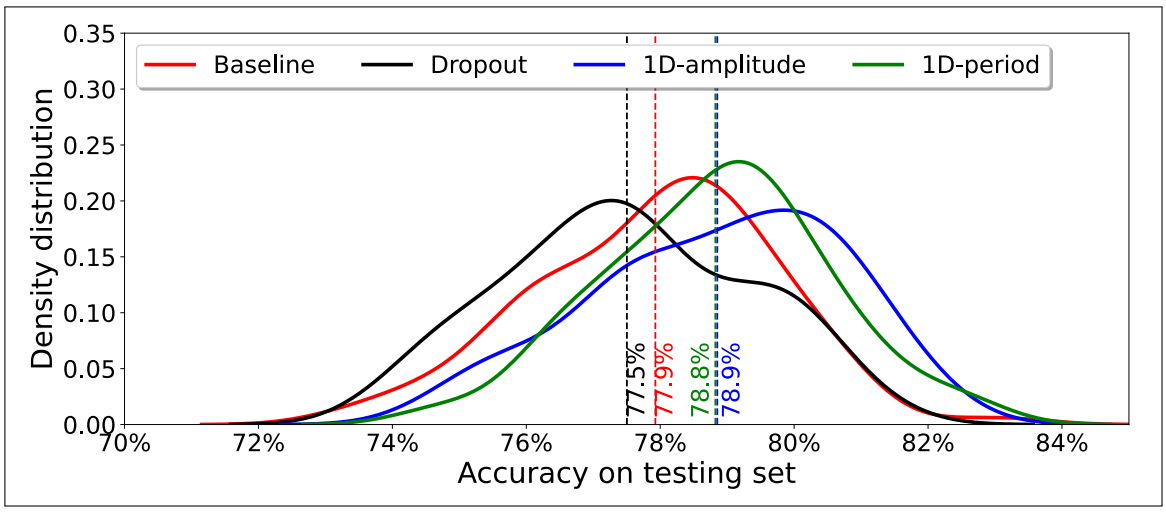


Figure 4.5. **Density distribution for the ACC in the testing set.** vertical lines show mean ACC for each training execution. Thirty experiments for each alternative were conducted, considering 5,000, 10,000 and 50,000 training cases. The x-axis is truncated to have a range from 70% to 85%.

indicate the first and third quartiles. The grey boxes show the standard deviation, and the white boxes contain the mean. Thirty experiments for each plot were carried out.

Figure 4.6(a) provides results for 1D-period signals, and Figure 4.6(b) for 1D-amplitude signals. When looking at Figure 4.6(a), we can observe that when the median (white boxes) is considered, our 1D-period signal approach always outperforms the baseline model. Similarly, Figure 4.6(b) presents the results for 1D-amplitude signals, where an improvement is also generated. The biggest improvement (1.8%) was obtained using 1D-period considering 50,000 training cases, being statistically significant according to the t-Student's test.

4.5.2. Regularisation using bidimensional knowledge injection

Figure 4.7 and Table 4.3 show the performance of our study's informative regularisation when using 2D-signals. Figure 4.7 provides violin plots, comparing ACC metrics in the testing data set, whereas Table 4.3 gives a summary of ACC, F_1-score and AUC metrics.

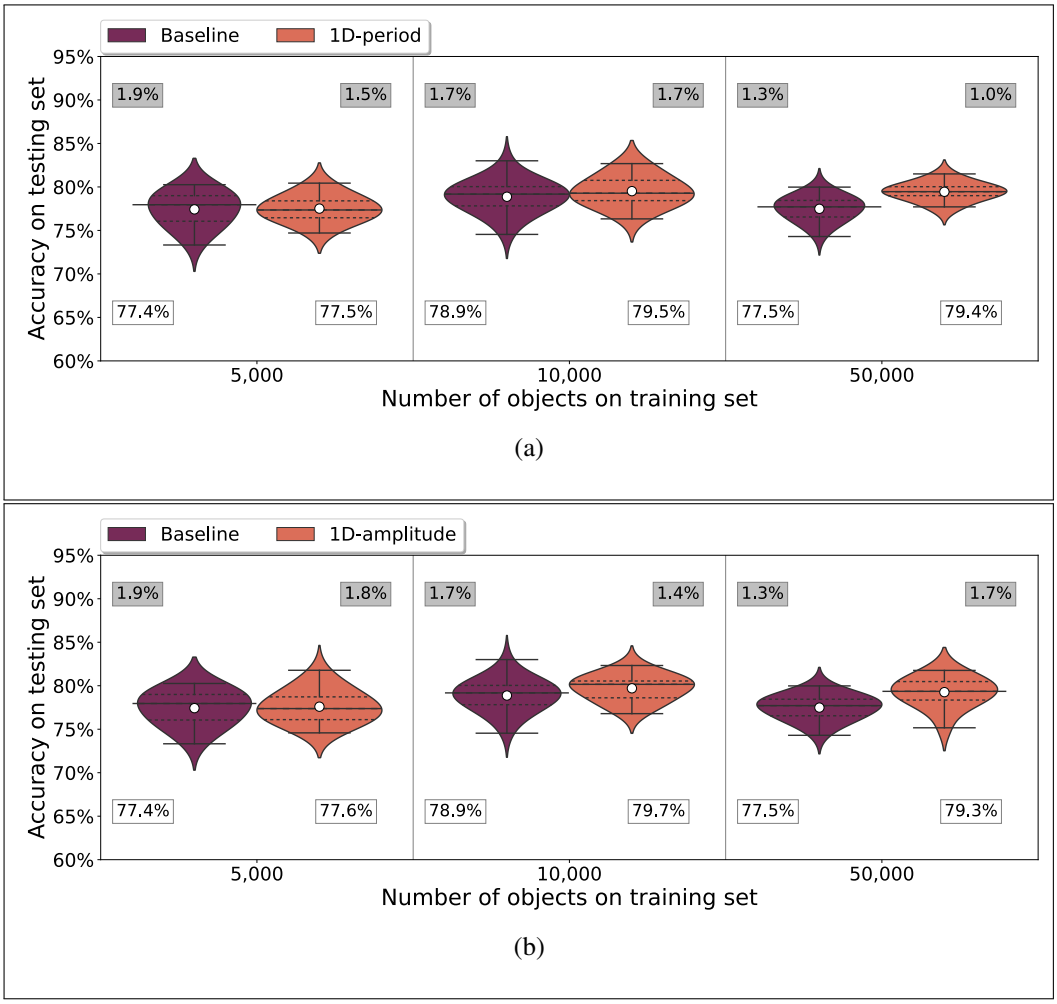


Figure 4.6. **Impact of regularisation using 1D signals on the ACC distribution.** White points represent the mean, black lines show the median, and dotted lines indicate the first and third quartiles. Grey boxes show the standard deviation value, whereas white boxes contain the mean value. Thirty experiments for each plot were undertaken. (a) provides results for 1D-period signals and (b) for 1D-amplitude signals. The y-axis is truncated to have a range from 60% to 95%.

We compare eight models, which are grouped as follows: (i) an MLP without regularisation, which is the baseline model; (ii) the baseline model using dropout as regularisation; (iii-iv) two norm-based regularisation approaches over the same model (l_1-norm and l_2-norm); and (v-viii) our four informative regularisation approaches using our training procedure. 2D-Gaussian denotes signals based on bivariate Gaussian distributions and 2D-uniform denotes 2D signals from uniform distributions (see Section 4.2).

When comparing the ACC metric in 1D signals (see Figure 4.6) with respect to 2D signals, we can highlight a significant improvement in the models trained using 50,000 light curves. In this case, 2D-Gaussian signals improve the ACC metric by 1.1% from 1D-amplitude signals and 0.9% from 1D-period signals. In the other cases (i.e., 5,000 and 10,000 training cases), 2D signals also improve the average ACC with respect to 1D signals. Regarding standard deviation, there is no significant difference between the baseline model and models regularised by signals.

The results from Table 4.3 show a significant improvement in the classification performance. The informative approaches improve the baseline model for the three settings ($n =$ 5,000, 10,000 and 50,000). Each metric is estimated on a set of thirty experiments. The impact of our regularisation proposal is noticeable, achieving the maximum increment in ACC metric ($\sim$3.0%) when the maximum result for the AUC metric is considered.

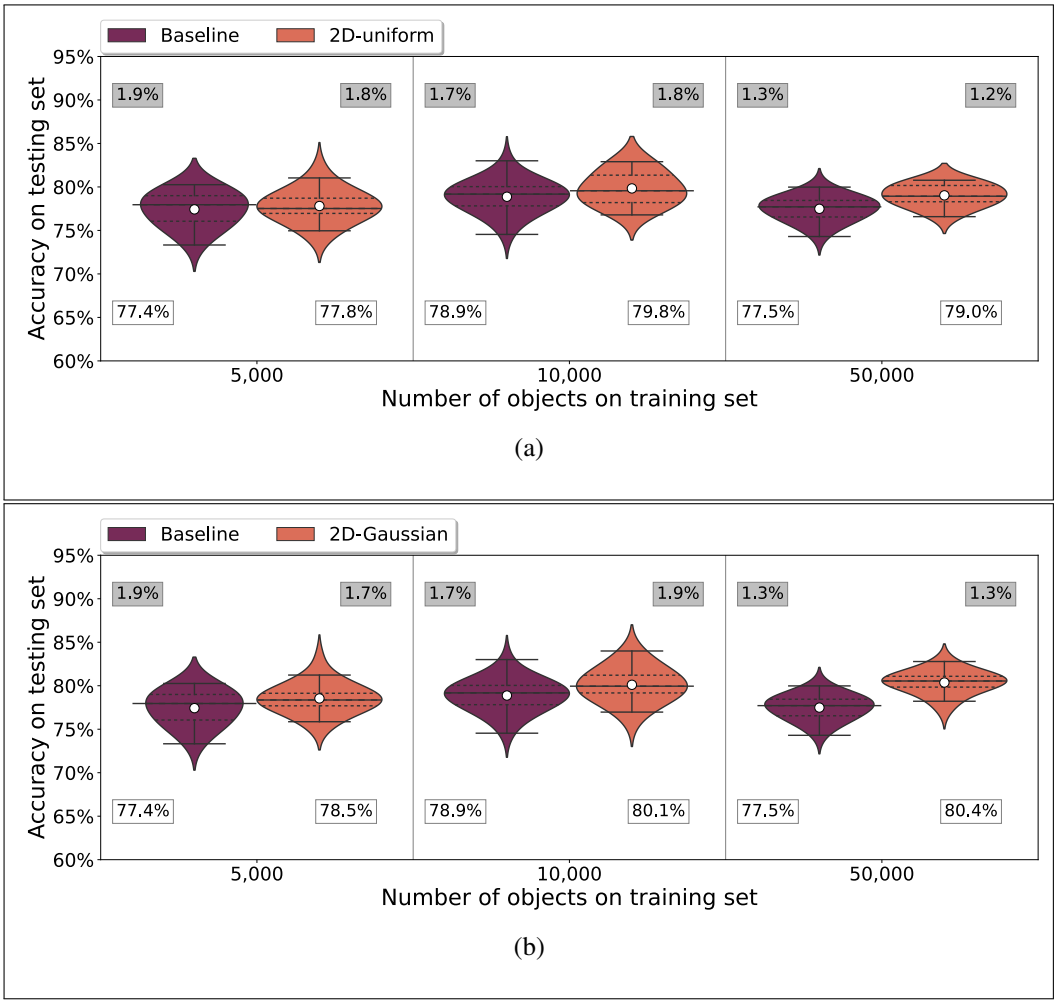


Figure 4.7. **Results of ACC in baseline model and informative regularisation using 2D signals.** (a) provides results for 2D signals from uniform signals and (b) for 2D signals using Gaussian signals. The number of experiments and information in the boxes is the same as in Figure 4.6. The y-axis is truncated to have a range from 60% to 95%.

Table 4.3. **Summary of results for three metrics (ACC, F_1-score and AUC) for the baseline model and informative regularisation.** The number of experiments is the same as in Figure 4.6. Early stopping is applied in all these experiments. The bold numbers represent the best strategy for model selection by each metric. * in bold numbers indicates a statistically significant difference in the baseline model. The t-student test was used to validate a difference in means.

n	Regularisation	ACC			F_1 score			AUC		
		max	median	mean	max	median	mean	max	median	mean
5,000	Baseline	80.26	77.96	77.42	79.98	77.40	77.42	0.87	0.84	0.84
	Dropout	80.03	76.80	76.76	79.74	77.24	77.16	0.85	0.83	0.82
	l_1-norm ($\lambda = 0.001$)	78.41	71.58	71.57	75.55	72.65	72.29	0.84	0.77	0.77
	l_2-norm ($\lambda = 0.001$)	**84.76**	76.83	77.01	79.77	76.14	76.44	**0.88**	0.84	0.84
	1D-amplitude	81.78	77.37	77.59	**80.66**	78.27	78.23	**0.88**	**0.85**	**0.85***
	1D-period	80.44	77.35	77.51	80.08	78.12	77.80	0.87	**0.85**	**0.85**
	2D-Gaussian	83.20	**78.36**	**78.54***	80.18	77.77	77.82	**0.88**	**0.85**	**0.85***
	2D-uniform	82.26	77.52	77.81	80.62	**78.40**	**78.48***	**0.88**	**0.85**	**0.85***
10,000	Baseline	83.01	79.18	78.87	81.35	**79.22**	79.15	0.88	0.85	0.85
	Dropout	81.33	79.07	78.47	80.49	78.74	78.53	0.85	0.84	0.83
	l_1-norm ($\lambda = 0.001$)	73.93	71.87	72.00	75.47	73.29	73.12	0.80	0.78	0.78
	l_2-norm ($\lambda = 0.001$)	**84.58**	78.56	78.21	80.23	77.36	77.37	**0.89**	**0.86**	**0.86***
	1D-amplitude	82.33	**80.16**	79.70	80.32	78.73	78.67	0.87	0.85	0.85
	1D-period	82.68	79.28	79.52	80.78	78.70	78.55	0.87	0.85	0.85
	2D-Gaussian	83.99	79.95	**80.11***	**81.86**	**79.22**	**79.17**	**0.89**	**0.86**	**0.86***
	2D-uniform	82.90	79.56	79.83	80.94	78.50	78.53	0.88	0.85	0.85
50,000	Baseline	79.98	77.71	77.49	84.59	83.04	82.98	0.87	0.85	0.85
	Dropout	80.58	77.16	77.28	85.00	82.76	82.75	0.87	0.84	0.84
	l_1-norm ($\lambda = 0.001$)	72.64	70.91	70.91	77.52	75.94	75.94	0.80	0.77	0.77
	l_2-norm ($\lambda = 0.001$)	77.51	73.05	73.42	82.98	79.71	79.65	**0.90**	0.87	0.86
	1D-amplitude	81.76	79.35	79.26	**85.19**	**83.42**	83.36	0.88	0.86	0.86
	1D-period	81.50	79.46	79.44	84.53	83.26	83.33	0.89	0.87	0.87
	2D-Gaussian	**82.78**	**80.54**	**80.37***	84.68	83.41	**83.47***	**0.90**	**0.88**	**0.88***
	2D-uniform	80.77	78.94	79.03	84.71	83.27	83.13	0.87	0.86	0.86

Concerning the ACC metric, when looking at experiments with $n = 5{,}000$ in Table 4.3, from informative approaches, the best results were obtained by 2D-Gaussian signals; the rest of the informative regularisation approaches also outperform the baseline model when the mean ACC is considered. The improvements obtained from 2D Gaussian signals with respect to the baseline model are 2.94%, 0.40%, and 1.12% for maximum, median and mean ACC, respectively. Even more remarkable, the 2D Gaussian signals outperform all the other approaches, including l_1-norm, l_2-norm and dropout in mean and median ACC; 2D Gaussian approach is only outperformed by l_2-norm when the maximum ACC is considered. For the experiment using $n = 10{,}000$, the maximum improvements come from l_2-norm (max), 1D amplitude-based signals (mean) and 2D Gaussian signals (median), 4.5%, 0.98% and 1.24% for maximum, median and mean, respectively. Lastly, when 50,000 training cases are used, the biggest improvements are generated by the 2D-Gaussian signals, obtaining an increment of 2.80% in the max metric, an increment of 2.83% in the median metric and an increment of 2.88% in the mean metric.

Regarding the F_1-score, we also observe (see Table 4.3) an improvement when applying signal-based regularisation. The 1D-amplitude signals provide the best result (max) with $n = 5{,}000$. On the other hand, considering the mean and median metrics, 2D-uniform signals provide the best scores. For the case $n = 10{,}000$ the best performance is obtained when 2D-Gaussian signals are injected, improving max F_1-score by 0.51% and median F_1-score by 0.02%; the scenario is different for $n = 50{,}000$, 1D-amplitude signals achieve the best results in max and median, and 2D-Gaussian outperforms the other alternatives when the mean metric is prioritised. We highlight that by assessing the F_1 score, informative approaches outperform traditional non-informative approaches such as dropout, l_1-norm and l_2-norm.

When looking at the AUC metric, we note that 2D-Gaussian signals and l_2-norm obtain the best results. 2D-Gaussian signals obtain the best performance in mean AUC for the three n considered. For the case $n = 5{,}000$ the other informative approaches have similar

performance; for $n = 10{,}000$ the two most competitive regularisation methods are 2D-Gaussian and l_2-norm. Lastly, when $n = 50{,}000$ 2D-Gaussian signals outperform the other proposals.

These results prove that our approach is a competitive regularisation strategy whose main advantage is incorporating human knowledge by using simple synthetic data (signals). Regarding traditional regularisation schemes, Dropout and l_1-norm regularisation methods cannot improve results according to our experiments. In fact, in most cases, these regularisation methods worsen the baseline model results. l_2-norm was able to improve the baseline model and, in some metrics (max ACC and AUC), was competitive with our approach.

To summarise, from these empirical results, we demonstrate a positive impact due to the incorporation of signals together with an ad-hoc training procedure. Our approach improved the baseline model performance by considering several metrics (max, mean and median for ACC, F_1-score, and AUC). Moreover, the improvements were, in the majority of cases, statistically significant, and they were observed in all those metrics.

5. CHAPTER V. A SELF-REGULATED CONVOLUTIONAL NEURAL NETWORK FOR CLASSIFYING VARIABLE STARS

5.1. Method overview

This chapter explains our final proposed methodology to train a more reliable classifier under biased data conditions, leveraging recent advancements in synthetic data generation based on deep learning models. By integrating a generative model with a classifier in a cooperative framework, our approach dynamically enhances the learning process with synthetic examples in underrepresented areas. During classifier training, the synthetic light curves focus on mitigating biases and imbalance issues. These samples are obtained from the stellar physical parameter space, which includes effective temperature, period, metallicity, absolute magnitude, surface gravity, and radius. The physical parameters are then processed by a trained PELS-VAE, which generates synthetic light curves. We emphasize that sampling from the physical parameter space allows us to effectively manage over- and under-represented zones due to its lower dimensionality when generating new light curves.

Our method adjusts the classifier's training trajectory by incorporating new objects from the generative model. We propose five policies for defining the number of samples for each class, some of which aim to populate classes where confusion is most significant, as indicated by the confusion matrix during the current training epoch. A mask-based training scheme is included to prevent competition between real and synthetic data. We also present a set of experiments to evaluate classifier performance under variations in the signal-to-noise ratio, sequence length, and training set size, highlighting where synthetic samples are most relevant to reduce data shift and address class imbalance. Finally, we provide evidence that our synthetic light curves can help train more reliable classifiers and optimize hyperparameters.

5.2. Classifier model

In this work, we propose a classifier designed to process $(\Delta t_i, \Delta m_i)_{i=1}^{L}$ sequences, formatted as $2 \times L$ arrays, where L is the sequence length. The classifier's architecture, which is provided in Figure 5.1, incorporates blocks of 1D convolutional layers, with the number of these blocks being a tunable hyperparameter from two to four. Each block comprises a 1D convolutional layer, followed by batch normalisation and a ReLU activation function. Additionally, a max pooling operator is integrated into each block to reduce the dimensionality and extract the most salient features. The first convolutional layer is equipped with 16 filters, each with a kernel size of six and a stride of one. Subsequent layers maintain the same kernel size and stride but double the filters in each new convolutional layer. After the convolutional layers, a fully connected layer is applied, the size of which depends on the number of convolutional layers due to the varying dimensions of the flattened output. The fully connected layer has an output dimension of 200. The network culminates with a final fully connected layer, which maps to the number of classes. We initialise the weights of all layers using the Xavier method (Glorot and Bengio, 2010).

A well-known approach to regularise the training process is incorporating synthetic samples (Goodfellow et al., 2016); however, some concerns must be considered to optimise the NN weights correctly. We highlight the balancing issue, which arises when there is a mismatch in the proportion of real and synthetic light curves used during training. An excessive number of synthetic light curves might lead the model to learn idealised data and inherit biases from the generation process, while too few may fail to provide sufficient regularisation. The main issue is that the number of synthetic light curves depends on the number of real light curves, which limits scalability. Because of that, we consider the idea of training a dual CNN where some filters are focused on learning real patterns from training data and the other is focused on learning synthetic data patterns. In this way, the number of synthetic samples does not depend on the number of real light curves.

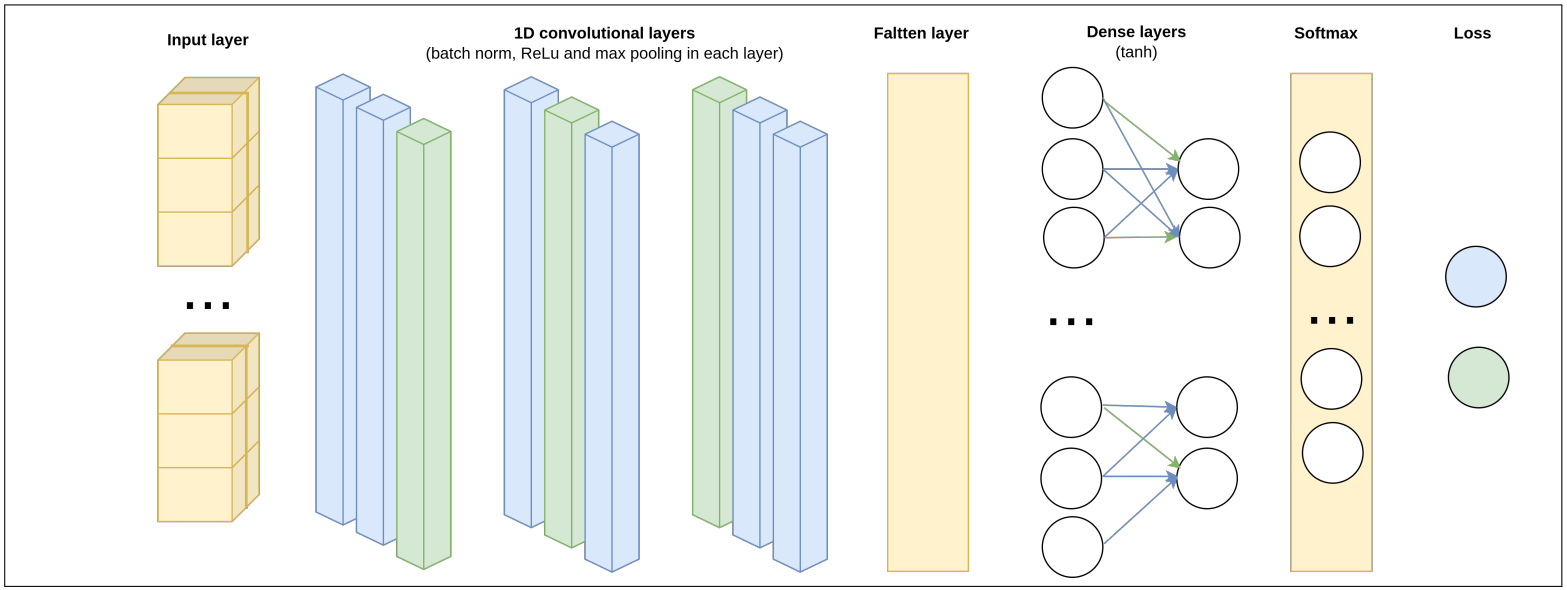


Figure 5.1. **CNN Classifier model architecture.** The input layer receives the $(\Delta t_i, \Delta m_i)_{i=1}^{L}$ representation. A sequence of 1D convolutional layers processes this input, where distinct sets of filters, managed by masks of weights, learn from real data ($\mathcal{L}_1$, depicted in light-blue) and synthetic data ($\mathcal{L}_2$, depicted in light-green), respectively. Each colour transformation corresponds to the filter operation it represents: light-blue for $\mathcal{L}_1$ and light-green for $\mathcal{L}_2$.

This mask-based training approach emulates two systems: a primary system that learns from training data and a secondary system that learns from synthetic data with well-known patterns. The secondary system acts as a regularizer, constraining the model's parameters by learning from synthetic data. This method was recently employed to train more reliable feature-based multi-layer perceptron (MLP) classifiers, successfully mitigating the data shift problem in MLP models (Pérez-Galarce et al., 2023).

Additionally, using a dedicated set of weights to learn synthetic light curves allows us to ensure the learning of the patterns injected synthetically. This approach helps to control the biases induced into the training set. We define weight masks as the selective activation of weights during training depending on the data type (real or synthetic), which is essential to our classifier architecture.

In Figure 5.1, each mask is represented by a colour: light blue for $mask_1$, which is designated for learning patterns from real light curves, and light green for $mask_2$, which focuses on synthetic light curves. In terms of architecture, the filters are assigned entirely to one mask based on a predefined quantile threshold in the convolutional and fully connected layers. Specifically, weights above this threshold are assigned to $mask_1$, tasked with learning from the training set. In contrast, weights at or below this threshold are assigned to $mask_2$, dedicated to learning from synthetic data. This segmentation of weights by quantile threshold is an inductive bias that regulates the model by simplifying its complexity. It guides the solution space (weights) towards zones likely to yield better generalisation, informed by the controlled introduction of synthetic samples.

This method ensures that each set of weights is optimised for its respective data type, potentially enhancing the ability of the model to discern and generalise from both real and synthetic patterns without overfitting. To control the trade-off between learning from training data and synthetic light curves, we define two hyperparameters, ε and ν. The parameter ε regulates the fraction of weights assigned to $mask_1$, while ν scales the learning rate for $mask_2$ relative to the learning rate of $mask_1$. For example, if ν is set to 0.5 and the learning rate for $mask_1$ is 0.1, then the learning rate for $mask_2$ will be 0.05. Setting these

parameters correctly, considering $mask_2$ as a regularisation, is crucial for good training convergence.

Due to the size difference between the training and the synthetic sets, the weights associated with $mask_1$ will updated more frequently than those in $mask_2$. To increase the utilisation of synthetic light curves during learning, we allow the repetition of the forward and back-propagation procedures within a single epoch for synthetic batches. The number of iterations is controlled by a hyperparameter which is optimised (`repetitions`).

5.2.1. Loss functions

Our architecture considers a different loss function for each mask of weights, i.e., for each type of light curve; therefore, in case of including parameters to control imbalanced class problems, each function can be customised. Three loss functions were proposed to assess the classifier and training robustness: cross-entropy, weighted cross-entropy and focal loss.

Cross-entropy: The cross-entropy loss function increases as the predicted probability diverges from the observed label. Thus, a perfect model would have a cross-entropy loss function of 0. The equation for a multi-class classification problem is given by

$$L_{\text{CE}} = -\sum_{i=1}^{C} y_i \log(p_i),$$

where C is the number of classes, y_i is a binary indicator (0 or 1) if class label i is the correct classification for the observation, and p_i is the predicted probability of the observation being of class i.

Weighted cross-entropy: A variant of the standard cross-entropy loss is useful in managing imbalanced datasets. This loss function is adapted by incorporating a factor that scales the importance of each class. For example, if a class is rare, a higher weight is assigned, making the model more sensitive to errors in classifying this class; this loss function is

expressed as

$$L_{\text{WCE}} = -\sum_{i=1}^{C} w_i y_i \log(p_i),$$

where, w_i represents the weight assigned to the i^{th} class. This weight is typically estimated as inversely proportional to the class frequency, giving more importance to less frequent classes. This modification can lead to some convergence problems when the classes are highly imbalanced. This potential issue encourages us to assess our training procedure using this loss function.

Focal loss: A loss designed to handle the issue of class imbalance by down-weighting well-classified examples. Its equation is given by

$$L_{\text{Focal}} = -\sum_{i=1}^{C} \alpha_i (1 - p_i)^{\gamma} y_i \log(p_i),$$

where α_i is a weighting factor for class i, $(1 - p_i)^{\gamma}$ is a modulating factor with tunable focusing parameter γ. The term $(1-p_i)^{\gamma}$ reduces the loss contribution from easy examples (where p_i is high) and increases the contribution from hard, misclassified examples.

5.2.2. Performance metrics

F_1: F_1 balances precision and recall by computing their harmonic mean. We utilise the macro-averaged F_1, which calculates F_1 for each class individually and then averages them, considering all classes equally. This is useful for assessing models in imbalanced datasets.

ROC-OVA: The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) evaluates the ability of a model to distinguish between classes. The one-vs-all approach computes the AUC-ROC for each class against all others, averaging by using the macro method, making it suitable for assessing multi-class classification. Its advantage lies in its simplicity and effectiveness in assessing each class independently, clearly understanding

how well each class is differentiated. In this case, we calculate the macro average over five ROC curves, one for each star type.

ROC-OVO: The one-vs-one AUC-ROC method takes a different approach. It compares every pair of classes, creating binary classifiers for each pair. The final AUC-ROC is the macro average of these comparisons, providing a detailed performance evaluation by considering the separability between all class pairs. By examining the separability of each class pair, this approach offers a comprehensive view of the model performance. ROC-OVO requires an average of $C(C-1)/2$ ROC curves, C is the number of classes; it is to say, one for each possible pair of type of star, which amounts to ten curves in our case.

5.3. Generative model

The generative model proposed by Martínez-Palomera et al. (2022) is a conditional variational autoencoder (VAE) modified to inject knowledge about physical parameters using temporal convolutional blocks and reconstructing folded and normalised light curves. While the original PELS-VAE was trained with *Gaia* DR2, in this work, we train the model using data from the *Gaia* Data Release 3 (DR3; Creevey et al., 2023), which allows us to include additional physical parameters. This enhancement encourages us to incorporate six key physical parameters: period, absolute magnitude, effective temperature, radius, surface gravity, and metallicity.

VAEs are generative models designed to model complex data distributions (Kingma and Welling, 2013). Given a dataset $X = \{x_1, x_2, \ldots, x_N\}$, a VAE defines a joint distribution over observed variables x and latent variables z as $p(x, z) = p(x|z)p(z)$. The objective function, often referred to as the evidence lower bound (ELBO), is given by

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{\mathrm{KL}}(q_\phi(z|x)||p(z)), \tag{5.1}$$

where θ and ϕ are the parameters for the decoder and encoder networks, respectively, $q_\phi(z|x)$ represents the approximate posterior distribution over the latent variables given

the data and D_{KL} is the Kullback-Leibler divergence. Traditional VAE has an important drawback known as the posterior collapse problem; this phenomenon implies that the VAE model ignores the latent variables and relies solely on the decoder (Lucas et al., 2019). A variant that allows managing this problem, improving, or disentangling the latent representation inducing a the balance between reconstruction quality and regularisation is the β-VAE (Higgins et al., 2017; Burgess et al., 2018); in β-VAE model, a new parameter β controls the relevance of each component in the loss function as follows,

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta D_{\mathrm{KL}}(q_\phi(z|x)||p(z)). \tag{5.2}$$

Despite the β-VAE model providing a more disentangled latent space, it is difficult to control the posterior distribution, for example, to explore how to change the posterior of a given light curve when we modify some physical parameters. To simplify this exploration was proposed the conditional variational autoencoder (cVAE), which extends VAEs by conditioning the model on additional information c; typically, the labels (Richards and Groener, 2022). In cVAEs, the joint distribution is modelled as $p(x, z|c) = p(x|z, c)p(z|c)$. The conditional ELBO becomes:

$$\begin{aligned}\mathcal{L}\Big(\theta, \phi; x, c\Big) = \mathbb{E}_{q_\phi(z|x,c)}\Big[\log p_\theta(x|z, c)\Big] \\ - \beta D_{\mathrm{KL}}\Big(q_\phi(z|x, c)\|p(z|c)\Big).\end{aligned} \tag{5.3}$$

Furthermore, the PELS-VAE considers injecting physical parameters. Temporal convolutional and dense layers are utilised in both the encoder and decoder. In the encoder, the physical parameters are incorporated after the temporal convolutional block and before the dense layers. It is important to note that this model reconstructs normalised and folded light curves; hence, additional procedures are required to use the synthetic light curves created by PELS-VAE in a $(\Delta t, \Delta m)_{n=1}^{L}$ input representation as required by our classifier. Several variants were tested to train this model, e.g. using imputation techniques to use the data from *Gaia* DR3 more efficiently, applying log transformation to the data sets, modifying β, etc.

5.4. Training process

Figure 5.2 provides an overview of one training epoch of our proposed approach. A condition is checked in each epoch to decide if a new synthetic set of samples from the generative model is required (step 1.1). If a new sample is required, a sampling strategy is applied to obtain synthetic physical parameters from a Bayesian Gaussian mixture model (step 2.1). After that, using a multiple outputs regression, the latent space is predicted (step 2.2). Then, using this latent space, the physical parameters, and the label, we obtain a folded and normalised light curve via a trained PELS-VAE (step 2.3). Finally, the folded and normalised light curves are transformed to obtain the representation used for the classifier model (step 2.4). In case no new synthetic samples are needed, first, an epoch is applied using the current synthetic light curves, updating only the weights belonging to $mask_2$ (step 1.2). Afterwards, another epoch is applied using the training set with real light curves, with updates limited to weights belonging to $mask_1$ (step 1.3).

5.4.1. Triggers of synthetic samples (step 1.1)

At each epoch, the training procedure assesses whether there is a need to generate a new batch of synthetic samples. This decision is based on two conditions. The first condition, controlled by the hyperparameter E, relates to the number of epochs elapsed without introducing new synthetic samples; it was set to three in the final experiments. The second condition, dictated by the hyperparameter ϕ, hinges on CNN's current proficiency in classifying synthetic samples. The hyperparameter E ensures a consistent introduction of diverse synthetic samples throughout the training process. Moreover, it is designed to gradually increase the generation of less probable light curves (reducing biases) as the training progresses. The hyperparameter ϕ, on the other hand, aims to increase the diversity and complexity of the synthetic data. It triggers the generation of new synthetic samples based on their classification difficulty, as indicated by the performance metric. Specifically, if the performance metric in classifying synthetic samples is excessively high, suggesting that they are too simple for the current model, it signals a need

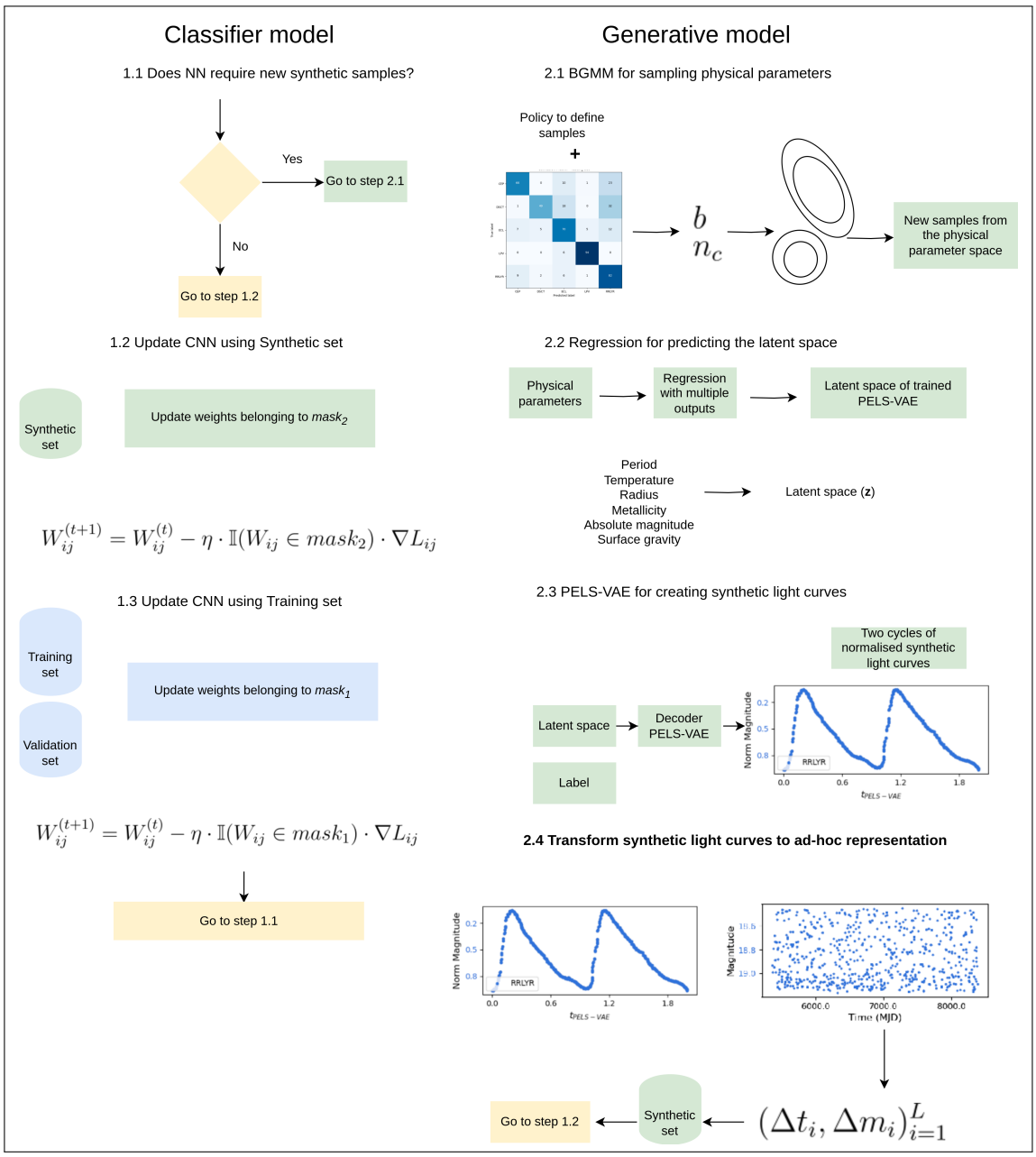


Figure 5.2. **Method overview of self-regulated training.** The process involves conditional checks for generating new synthetic samples, sampling strategies for obtaining synthetic physical parameters, multiple outputs regression for predicting latent space, and transformations for obtaining representations used by the classifier model.

for more challenging samples. This hyperparameter was optimised in the hyperparameter search.

In addition, we need to decide how many objects will be sampled for each class. We designed the following policies based on the confusion matrix to define the number of samples from each variable star type.

`Correct classification rate` (`CCR`): This policy calculates the `CCR` for each class, which is the ratio of correct predictions to the total number of instances for that class. It ranks the classes based on these ratios in ascending order, considering a factor that reduces the number of objects according to the ranking. For example, if a factor is equal to two, the star class in position $k + 1$ will have half the objects with respect to position k.

`Max confusion`: This approach focuses on the classes most confused with others. Given a confusion matrix, it sums the off-diagonal elements in each row and ranks the classes based on these sums. This policy uses the same criteria as `CCR` to assign the number of objects for each variable star type.

`max_pairwise_confusion`: This procedure iteratively finds pairs of classes with maximum confusion and ranks the classes based on these pairs. This policy uses the same criteria as `CCR` to assign the number of objects for each variable star type.

`Proportion`: This method normalises the confusion matrix, considering off-diagonal elements, and then assigns a proportional budget of samples to each cell.

`non-priority`: Equal quantity of samples for each star type.

5.4.2. Sampling method (step 2.1)

The first step in the synthetic light curve generation involves sampling from the physical parameter space, and this sample is then fed into the PELS-VAE. The number of objects per class, n_c, is defined according to previously described policies, with one of these objects being used for the complete training. To generate these n_c physical parameter samples, we use a Bayesian Gaussian mixture model (BGMM) for each variable star

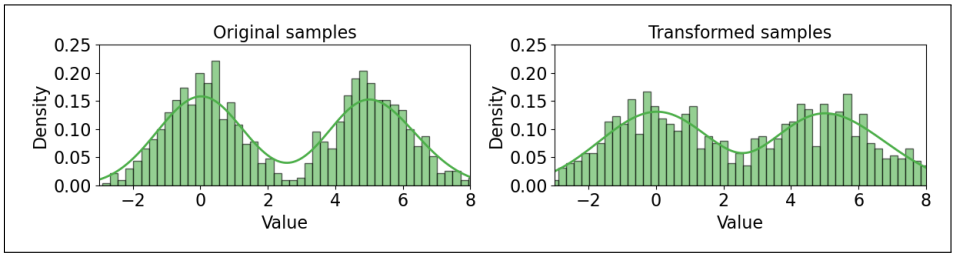


Figure 5.3. **Comparison of original and transformed samples from a GMM.** This random variable (Value) does not have physical meaning, it is only used to illustrate the distribution modification. The histogram on the left shows the distribution of samples from the original distribution ($\mathcal{N}\left(\begin{bmatrix}0\\5\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$), and the histogram on the right shows the modified distribution using a $b = 0.6$, which accentuates the tails of the original distribution. Section 5.4.2 dives into this distribution adaptation.

type. The reasons for selecting a BGMM include: (i) it allows us to inject expert knowledge about the means directly, preventing unstable behaviour when the class contains few examples; (ii) it eliminates the need to determine the optimal number of subpopulations; and (iii) each BGMM enables us to quantify zones that are more or less probable for physical parameters for each class. The means for each star type are also obtained from the expert knowledge (see Appendix C.2). To fit the BGMM, we use the variational inference (Blei and Jordan, 2006) approach implemented in `sklearn` (Pedregosa et al., 2011).

To control the biases, we define a modified probability density distribution q based on the target probability density distribution p as $q(x) = \frac{p(x)^b}{Z}$. Here, Z represents a normalisation constant, calculated as $Z = \int p(x)^b \, dx$. While direct evaluation of this integral might be computationally daunting, its exact value often becomes irrelevant in the sampling framework. The parameter b plays a pivotal role in the modified density. When $b = 1$, q mirrors p, resulting in a traditional sampling process; if $b > 1$, the modified density accentuates differences, focusing sampling around p's modes; and finally, when $0 < b < 1$, q explores regions less represented in p. Figure 5.3 shows a toy one-dimensional example of how sample distribution is modified by b, where $p(x)_{\text{new}} = \dfrac{p(x)^b}{Z}$.

This sampling method is designed to generate synthetic light curves by varying physical parameters, enabling efficient exploration of the physical parameter space. The primary aim is to prioritise and represent less explored zones in the training dataset. The

hyperparameter b plays a crucial role in this process by adjusting the density distribution of these physical parameters. Specifically, when the value of b is reduced, it effectively diminishes the peaks of the modes in the distribution. This modification ensures a more diverse and representative parameter space sampling, generating a wider variety of synthetic light curves. An exponential decay function models the decay of b, $b = b_f + (1 - b_f)\exp(-c \times \text{epoch})$, where two additional hyperparameters, b_f and c, must be defined. Intuitively, these hyperparameters model the minimum b value by b_f, which sets the maximal modification of the initial probability density distribution and the speed of this decay process, controlled by c.

5.4.3. Regression to latent space (step 2.2)

The training of our regression model adheres to the conceptual framework outlined by Martínez-Palomera et al. (2022). Specifically, an RF model, known for its capability to handle multiple outputs, is employed. The input features for this model are the physical parameters extracted from *Gaia* DR3. The regression's target variables are the latent space representations of stars obtained by processing each star's data through the trained PELS-VAE. A grid search with k-fold cross-validation is utilised to optimise the hyperparameters (number of estimators, max depth, minimum number of samples in each split and minimum number of samples in each leaf). For performance evaluation, 20% of the data is reserved for testing, and the mean absolute error is employed as the error metric. Once the RF regressor is trained, it becomes integral to the process immediately following the generation of synthetic samples. Its primary role is to predict the latent space based on a given set of physical parameters.

5.4.4. Generate samples (Step 2.3)

Once the latent space is established, the decoder component of the PELS-VAE becomes essential for generating new synthetic light curves. We can generate a normalised

light curve by inputting to the PELS_VAE decoder a specific class and the corresponding latent space, which is obtained from physical parameters.

5.4.5. Adapt representation and create synthetic batch (Step 2.4)

The final stage of synthetic batch generation consists of three key steps: magnitude scaling, temporal sampling, and difference computation. During magnitude scaling, we select a real light curve from a star of the same type as those in the synthetic sample, choosing the one with the closest matching period. This real light curve is used to determine the maximum and minimum magnitudes, which are then applied to scale the synthetic light curve using the min-max scaling method as follows:

$$m' = m_{\mathrm{PELS-VAE}}(m_{\max} - m_{\min}) + m_{\min} + \varepsilon,$$

where $m_{\max}$ and $m_{\min}$ are the maximum and minimum magnitudes from the real light curve, $m_{\mathrm{PELS-VAE}}$ is the normalized magnitude output by the PELS-VAE, and $\varepsilon \sim \mathcal{N}(\mu_p, \sigma_p^2)$ is a noise term, where μ_p and σ_p^2 represent the mean and variance of the photometric error.

Temporal sampling is performed by using the sampled period and calculating the number of cycles based on the baseline of the real light curve. The adjusted time, denoted as t', is computed as:

$$t' = t_{\min} + k \times t_{\mathrm{PELS-VAE}},$$

where k represents the number of cycles, $t_{\min}$ is the minimum time recorded in the real light curve and $t_{\mathrm{PELS-VAE}}$ is the temporal dimension which is generated by the PELS-VAE. For each observation, k is sampled from a uniform distribution with bounds $a = 0$ and $b =$ the maximum number of cycles in the real light curve mentioned earlier.

Finally, we sort the observations in temporal order and compute the time and magnitude differences for each synthetic light curve, thus completing the synthetic batch generation process. Figure 5.5 illustrates this process, from the PELS-VAE output to the CNN input.

It is important to note that the PELS-VAE model is capable of generating sequences that are longer than those in the training set. Consequently, it is possible to sample multiple sequences from each synthetic light curve. The number of samples per synthetic light curve is governed by a hyperparameter, `n_oversampling`. The choice of sampling method is particularly crucial for short sequences. If a random sampling approach is applied, there is a risk of omitting significant portions of the phase. In such cases, an equally spaced sampling strategy can be employed, with adjustments made to the first and last observations.

5.5. Hyperparameters

To manage this complex hyperparameter optimisation, we utilised a Bayesian optimisation framework implemented through `Weights & Biases` to tune the hyperparameters (Biewald, 2020). Two metrics were evaluated to identify the optimal parameters: the F_1^{val} score and a weighted F_1 score, defined as:

$$\mathrm{F}_1^{\text{weighted},\alpha} = (1-\alpha) * \mathrm{F}_1^{\text{val}} + \alpha * \mathrm{F}_1^{\text{synthetic}}.$$

$\mathrm{F}_1^{\text{weighted},\alpha}$ score aims to leverage synthetic light curves to guide the exploration of the hyperparameter space, serving as a second layer in directing the learning process. In other words, we are applying regularisation to both the learning parameters (weights) and the hyperparameter optimisation. We selected macro weighting for each F_1 score since it is more suitable for imbalanced classes; thus, during hyperparameter exploration, the search is directed toward models with better performance in scenarios with imbalanced classes. Given the extensive list of parameters and their range of alternatives, this hyperparameter search was supplemented with a grid-search exploration in preliminary stages, allowing for a better balance of exploration and intensification, i.e., exploration extensively searches the solution space to avoid local optima while intensification exhaustively searches promising regions to find the best solution (**?**).

5.6. Data

5.6.1. OGLE

The classification approach was tested using the OGLE-III online catalogue of variable stars[1], which is part of the Optical Gravitational Lensing Experiment (OGLE; Udalski et al., 2008). The OGLE project is a long-term project aimed at detecting microlensing events and identifying exoplanets via transit methods, including observations from critical astronomical areas such as the Galactic bulge, the Large and Small Magellanic Clouds, and the Carina constellation. To train and test our model, we utilised 419,257 light curves from the I band, conducting various experiments to assess the impact of different, training set size, signal-to-noise ratio levels and sequence lengths.

5.6.2. *Gaia* DR3

In the training of our PELS-VAE model, we utilise six key physical parameters derived from the *Gaia* DR3 catalogue, one of the most comprehensive resources for astrophysical data available (Creevey et al., 2023). The parameters considered are metallicity ([Fe/H], measured in dex), period (P, in days), absolute magnitude in the G band (M_G), surface gravity ($\log g$, in dex), radius (R, in $R_\odot$), and effective temperature ($T_{\rm eff}$, in K). While we specifically use the $[\mathrm{Fe/H}]_{\mathrm{J95}}$ metallicity scale (Jurcsik, 1995), we refer to it simply as [Fe/H] in subsequent sections. We applied logarithmic transformations to the effective temperature, period, and radius to scale the data distribution. Figure 5.4 shows the distribution of these parameters.

To integrate observational data from *Gaia* DR3 with OGLE, we utilised the cross-matching technique described by Martínez-Palomera et al. (2022), successfully matching 53,090 stars. We decided to impute physical parameters after preliminary experiments during the PELS-VAE training with and without imputation. Imputation was performed stratified by class using the k-nearest neighbour (KNN) method (Troyanskaya et al., 2001)

[1] https://ogledb.astrouw.edu.pl/ ogle/CVS/

with $k = 5$. This method not only allows us to use a bigger training set but also ensures that imputed values represent the nearest neighbours within the same class, avoiding confusion with objects from different classes and similar physical parameters. Table C.1 details the missing data within the physical parameters.

When examining Figure 5.4, which shows the physical parameters used to enhance the PELS-VAE, it becomes clear that the stellar classes exhibit distinct characteristics in these physical parameters that encourage us to start the synthetic curve generation in this space. The univariate period distribution is particularly informative; four stellar classes can be separable. Eclipsing binaries tend to overlap with other categories, which could indicate similarities in their physical parameters. Specifically, the delta Scuti stars (DSCT) are noticeable in the lower period range. Following these, the RR Lyrae stars are observable, with the Cepheid stars and then the long period variables being identifiable at progressively higher periods. In the bivariate log radius versus log period plot, a well-known pattern appears where the Cepheid stars display a pronounced positive correlation, showcasing a direct proportionality between their pulsation periods and radius. This relationship is consistent with the established period-luminosity relation. When we delve into the multivariate space – considering other parameters such as metallicity, effective temperature, absolute G magnitude, and surface gravity – the separability of these classes does not always become more pronounced. Long-period variable (LPV) stars are separable for all these physical parameters in this set. The data depicted in Figure 5.4 can be complemented by the traditional stellar classifications of the Hertzsprung-Russell diagram by showing variable stars with diverse periods, radii, and luminosities. It confirms the established correlations, such as the period-luminosity relationship for Cepheids.

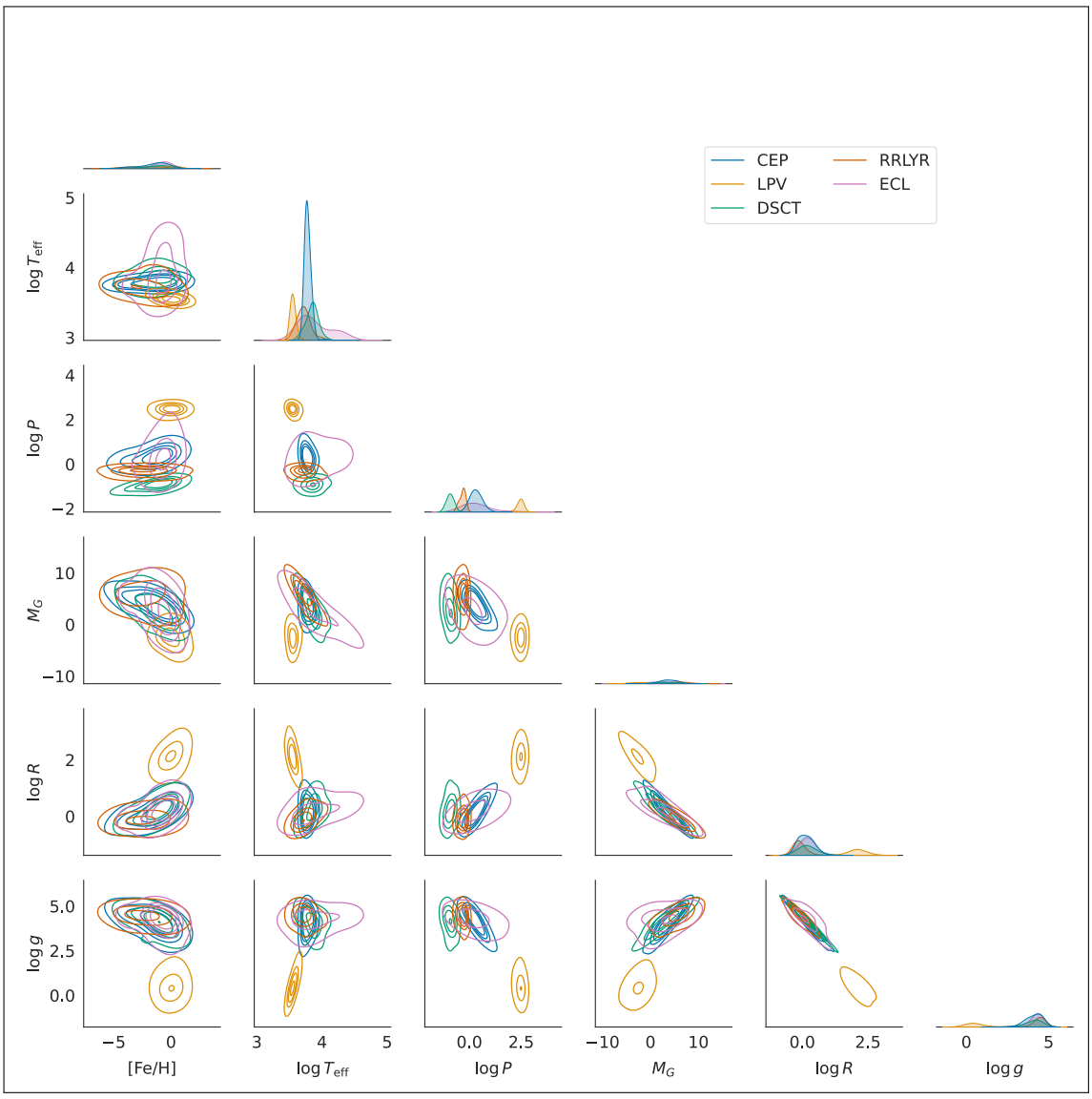


Figure 5.4. Physical parameters by class extracted from *Gaia* DR3.

5.6.3. Induced biases

Two data shift scenarios were designed to validate our proposal. These benchmark sets aim to replicate the inherent biases present in the training data of variable stars. To generate these benchmark sets, we employ an active learning metric based on uncertainty sampling; specifically, we utilise the Gini impurity index ($G_{\rm index}$), which is defined as $G_{\rm index} = 1 - \sum_{k=1}^{|K|} p_k^2$, where $|K|$ represents the number of classes and p_k is the probability of selecting an element from class k. For a given instance, the $G_{\rm index}$ quantifies how mixed the class probabilities are. If an instance has a high $G_{\rm index}$, the model is more uncertain about which class the instance belongs to.

To obtain p_k, we first have to fit a probabilistic classifier; it has to be trained on the entire catalogue to generate soft predictions for each star, representing the probability of belonging to a particular class. Using these probabilistic outcomes, we can quantify the uncertainty level for each star, which can be interpreted in our context as the classification difficulty. To facilitate the allocation of objects to the training or testing sets without employing a strict threshold (i.e., if $G_{\rm index} \leq$ threshold $\rightarrow$ train else testing), we incorporate randomness into the selection process, modulated by a tempering constant T as follows. For each star, we first compute the probability $p = e^{-G_{\rm index}/T}$ and then sample a random value r from a uniform distribution between 0 and 1. If $r \leq p$, the data point is added to the training dataset; otherwise, it is added to the testing dataset. In this way, uncertain objects are more likely to be included in the testing set. The parameter T controls the degree of shift in the dataset, with higher T values yielding more balanced sets.

The first induced bias, which generates *Data set I*, involves training a binary RF classifier for RR Lyrae stars using features estimated through the FATS package (Nun et al., 2015). In this case, the bias is introduced by focusing on including stars that are more likely to be uncertain and could be confused with RR Lyrae stars in the testing set; for this scenario, we used $T = 1$. The second case, which generates *Data set II*, considers an RF multi-class classifier. In this scenario, we used $T = 3$. The bias in this scenario

is designed to induce confusion among all classes by introducing ambiguity in class labels across the dataset. This approach helps assess how well the model can generalise when faced with data shifts involving multi-class overlapping in the testing set. Table 5.1 presents the distribution of objects within the testing and training sets for each induced bias.

We used the Mann-Whitney U test (MacFarland and Yates, 2016), a univariate non-parametric method, to statistically validate the differences in period and amplitude between the training and testing sets for both datasets and across all classes. Specifically, the null hypothesis (H_0) states that the two populations are identical. In the majority of cases, the results rejected the null hypothesis, indicating significant group differences; the only exception is for DSCT stars in the period feature, where the null hypothesis was not rejected. This result is likely influenced by the limited number of samples available for this star type. Given these results, we confirm that the induced biases within the benchmark data sets contain the data shift problem, making them suitable for testing our approach.

Table 5.1. **Number of objects and percentage per class in training and testing sets for *Data set I* and *Data set II*.** The validation set is obtained from the training set (30%) in a stratified sample.

Class	Class name	*Data set I*		*Data set II*	
		Training (occurrences, %)	Testing (occurrences, %)	Training (occurrences, %)	Testing (occurrences, %)
CEP	Cepheids	4,093 (1.05%)	3,859 (13.63%)	4,472 (1.14%)	3,480 (12.23%)
DSCT	Delta Scuti	1,496 (0.38%)	1,311 (4.63%)	1,447 (0.37%)	1,360 (4.78%)
ECL	Eclipsing Binaries	35,061 (8.97%)	6,726 (23.76%)	34,652 (8.87%)	7,135 (25.08%)
LPV	LPVs	321,451 (82.22%)	2,527 (8.93%)	316,092 (80.88%)	7,886 (27.72%)
RRLYR	RR Lyrae	28,847 (7.38%)	13,886 (49.05%)	34,144 (8.74%)	8,589 (30.19%)
Total		390,948	28,309	390,807	28,450

5.7. Results

5.7.1. Hyperparameter selection

Given that the data shift problem is crucial in our experimental setting, it is unclear if traditional performance metrics like F_1 or ACC on the validation set work well in exploring the hyperparameter space. Because of that, we compare a traditional approach using an F_1^{val} and $F_1^{\text{weighted},\alpha}$. In the $F_1^{\text{weighted},\alpha}$, synthetic light curves were considered during the hyperparameter exploration but were not used to calculate the score for selecting the best hyperparameter set.

Table 5.2 compares the three search methods across both benchmark sets, including the best F_1 test values found during exploration (BF), F_1 scores for recommended hyperparameters (RH), and the policy methods that achieved these values. The objective score that provided the best results in both sets was the weighted $F_1^{0.15}$, indicating that incorporating synthetic samples to define model performance can help to explore the hyperparameter space. The hyperparameter search yielded better results not only in the RH but also in the BF. This empirical evidence encourages paying more attention to hyperparameter optimisation in light of underlying data issues in variable stars, such as data shift or class imbalance problems.

Table 5.2. **Hyperparameter optimisation.** RH: objective score for recommended hyperparameters, BF: best found objective score, BFP: policy used in best-found solution, RFP: the policy used in the recommended solution.

Data set	Exploration	BF	RH	BFP	RFP
Data set I	F_1 val	0.56	0.52	`non-priority`	`max_pairwise_confusion`
	$F_1^{\text{weighted},0.15}$	**0.58**	**0.54**	`Max confusion`	`max_pairwise_confusion`
	$F_1^{\text{weighted},0.3}$	0.57	0.52	`Proportion`	`max_pairwise_confusion`
Data set II	F_1^{val}	0.65	0.65	`Proportion`	`Proportion`
	$F_1^{\text{weighted},0.15}$	**0.67**	**0.66**	`Max confusion`	`max_pairwise_confusion`
	$F_1^{\text{weighted},0.3}$	0.66	0.65	`Proportion`	`CCR`

5.7.2. Synthetic light curves

This section focuses on providing an example of the data pipeline from physical parameter samples to the input representation fed to the classifier. It is worth highlighting that this data pipeline can be used independently for another classifier as a tool for synthetic data generation; even using a different input representation, it only receives the b parameter as input, which controls the mode peaks when sampling physical parameters. Figure 5.5 shows the dataflow from light curves predicted by the PELS-VAE to the $(\Delta t, \Delta m)_{n=1}^{L}$ representation. The first column displays the synthetic light curves generated by the PELS-VAE. The second column presents the same light curves, converted into raw light curves as explained in Section 5.4.5. The third column shows the folded light curves, including photometric errors, alongside a real light curve from the training set with similar physical parameters, identified using KNN; this column is displayed solely to validate the conversions. Finally, the last column shows the $(\Delta t_i, \Delta m_i)_{i=1}^{L}$ representation, which is the input for our proposed classifier. Additional examples of synthetic light curves, compared with the closet object from the training set, are presented in Figure 5.6. It is important to note that replicating real light curves is not our objective. Instead, we are comparing the synthetic light curves with similar real light curves in the physical parameter space; in fact, the physical parameters used for the generated light curves do not exist within our training set. Moreover, replicating an existing light curve from the training set is simpler, as the model has already seen it and likely learned it perfectly. However, such light curves are not useful for addressing the data shift problem. Our primary focus is on generating synthetic light curves from under-represented regions of the physical parameter space, where data is scarce. Although the synthetic light curves are not perfect, the patterns for each star type are clear, which has enabled us to improve the classifier's reliability.

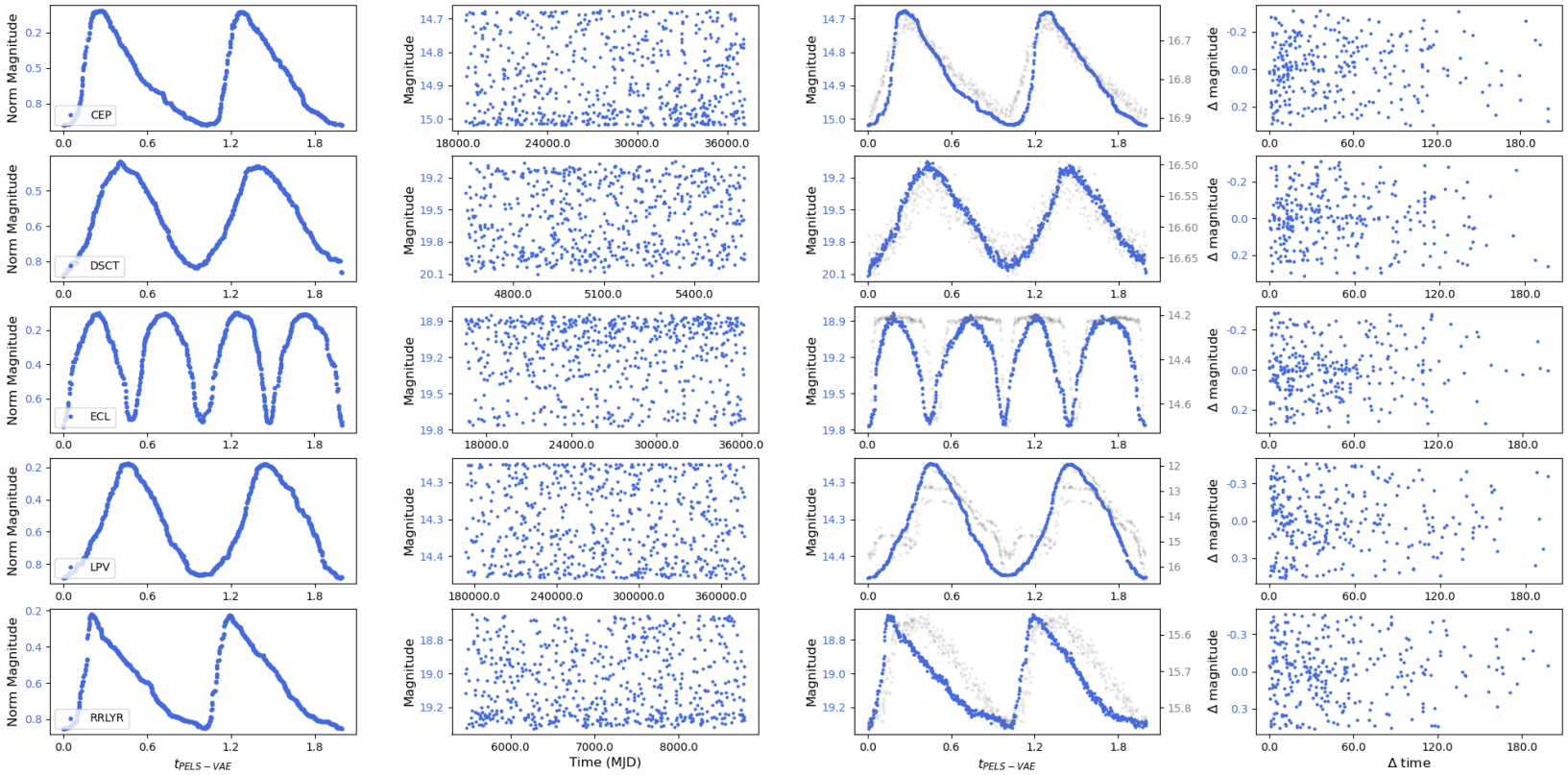


Figure 5.5. **Dataflow for synthetic light curves.** The first column shows the light curves predicted by the PELS-VAE for these samples, i.e., the normalised and phased light curve outputted from the PELS-VAE. The second column presents a reverted light curve according to the procedure explained in Section 5.4.5. The third column illustrates a synthetic folded light curve (blue) with respect to a light curve in the training set (grey) with similar physical parameters to validate the process and visualise the impact of added noise. Finally, the last column provides the $(\Delta t, \Delta m)_{n=1}^{L}$ representation. Each row corresponds to a different variable star type (Cepheids, delta Scuti, eclipsing binaries, LPV and RR Lyrae).

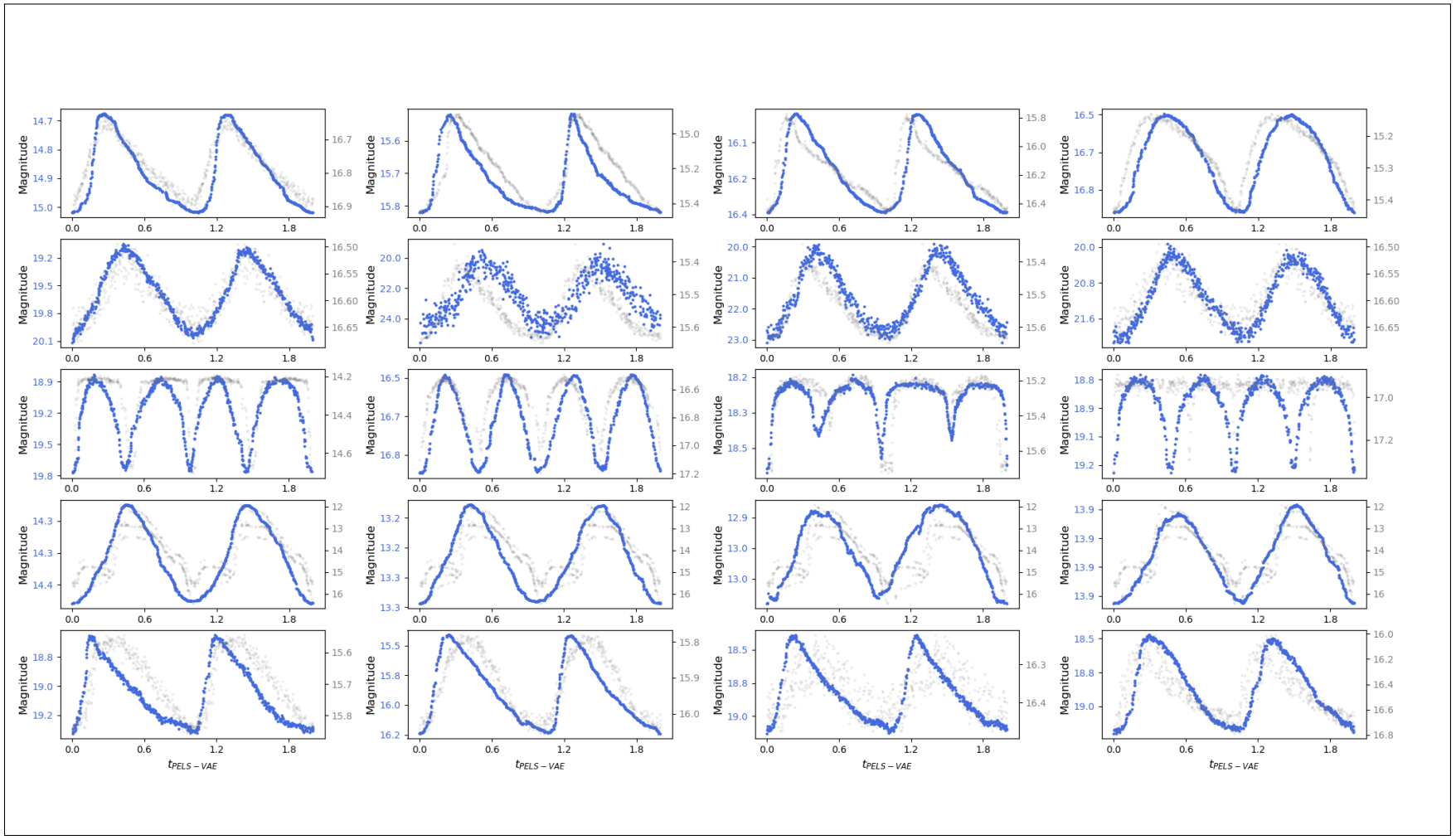


Figure 5.6. **Different samples of synthetic light curves compared with the closest real light curves.** They were matched using k-nearest neighbours (kNN) based on physical parameters. The first column shows the light curves predicted by the PELS-VAE for these samples, i.e., the normalised and phased light curve outputted from the PELS-VAE. The second column presents a reverted light curve according to the procedure explained in Section 5.4.5. The third column illustrates a synthetic folded light curve (blue) with respect to a light curve in the training set (grey) with similar physical parameters to validate the process and visualise the impact of added noise. Finally, the last column provides the $(\Delta t, \Delta m)_{n=1}^{L}$ representation. The star categories are ordered from top to bottom as follows: Cepheids, delta Scuti, eclipsing binaries, long period variables, and RR Lyrae stars.

5.7.3. Comparison of policies and loss functions

Table 5.3 presents the experimental results for different policies, including the mean, minimum, and maximum values for F_1, ROC-OVO, and ROC-OVA. For *Data set I*, the proportional policy outperforms the other policies across all metrics. In the case of *Data set II*, the `no_priority` policy achieves the highest value in seven of the nine metrics. The maximum values for the three metrics show higher variability, as the maximum value for F_1 was obtained by the `max_pairwise_confusion` policy, the maximum value for ROC-OVO was achieved by the `CCR` (and `no_priority`) policies, and the maximum value for ROC-OVA was found by the max_confusion policy. Upon examining Table 5.3, significant performance differences can be observed when different policies are used to assign the number of synthetic stars, underscoring the importance of policy selection. Moreover, the varying results across the two different data sets emphasise the need to choose policies based on the specific characteristics of each data set. The `proportion` policy and `no_priority` policy will be used in the following sections for experiments with *Data set I* and *Data set II*, respectively.

Table 5.4 shows results for different loss functions using our training approach on the *Data set I*. Despite some differences in these metrics, all loss functions perform well within this training approach. Cross entropy, which does not consider weights for managing the imbalance problem, exhibits higher variability in results, obtaining several worst performances, but it also achieves the best performance in other metrics. Focal loss and weighted cross entropy never obtain the worst performance and achieve some of the best performances. Weighted cross entropy achieved the best results in all metrics related to the F_1, while focal loss obtained the best results in both minimum values for the ROC-based metrics. In the following experiments, we will use the focal loss function due to its greater flexibility in the hyperparameter search and performance comparable to that of weighted cross-entropy.

Table 5.3. **Statistics for policies and sample size in testing sets with mean, minimum, and maximum values.** The highest values in each column are highlighted in bold, whereas the lowest values are coloured red. Ten independent samplings were evaluated in each row, and 40,000 light curves were considered in the training set for each experiment.

	Policy	F_1			ROC-OVO			ROC-OVA		
		mean	min	max	mean	min	max	mean	min	max
Data set I	`CCR`	56.4	53.7	59.3	88.0	85.3	89.6	87.0	85.4	88.7
	`max_confusion`	56.5	53.4	59.5	88.2	85.9	89.2	87.1	84.6	88.1
	`max_pairwise_confusion`	56.7	52.2	60.9	88.2	86.5	90.1	87.1	85.2	89.2
	`no_priority`	57.3	54.4	60.1	88.6	87.5	89.9	87.5	86.1	88.6
	`Proportion`	**57.8**	**54.6**	**61.9**	**88.8**	**87.7**	**90.6**	**87.8**	**86.9**	**89.6**
Data set II	`CCR`	57.5	52.2	61.4	88.9	85.1	**90.4**	89.9	86.6	91.2
	`max_confusion`	58.1	52.7	61.2	88.8	86.3	91.0	89.8	87.4	**91.7**
	`max_pairwise_confusion`	58.1	51.2	**62.9**	88.8	86.4	90.3	89.7	87.6	91.0
	`no_priority`	**59.0**	**54.5**	62.3	**89.2**	**88.0**	**90.4**	**90.0**	**89.1**	91.1
	`Proportion`	57.7	49.3	62.0	88.3	82.9	90.1	89.3	85.0	90.8

5.7.4. Signal-to-noise ratio and sequence length impact

When examining Table 5.5, we observe that incorporating synthetic samples and mask-based training (referred to as "Two losses" or self-regulated training) leads to improvements in all performance metrics with respect to the training without masks for regularisation (referred to as "One losses"). For *Data set I*, the F_1 score increases by a minimum of 1.6 and up to 3.5 on average, with a maximum improvement of 5.7 in the minimum F_1 column for an sn_ratio of 4. Regarding the ROC metrics, the mean increase in ROC-OVO ranges from 0.4 to 0.8, while the mean increase in ROC-OVA ranges from 0.6 to 1.0. Similarly, for *Data set II*, we observe that applying our training approach with synthetic samples leads to consistent improvements, particularly in the F_1 score, which increases by 1.4 on average when the sn_ratio is 4, with a maximum improvement of 8.1 for the minimum F_1 obtained from the experiments. The ROC metrics for *Data set II* show more modest gains; however, a clear trend of improvement is evident. Table 5.5 also shows that classification performance remains stable even as the signal-to-noise ratio decreases, indicating that our approach effectively handles high noise levels in observations without compromising performance metrics. Interestingly, in some cases, particularly for F_1 with self-regulation, the results for an sn_ratio of 4 outperform those for an sn_ratio of 6. This counterintuitive pattern may be linked to improved regularisation and calibration, achieved through the use of light curves with a lower sn_ratio in combination with synthetic light curves. Since F_1 is more sensitive to poor calibration than ROC-based metrics, it may benefit from this specific experimental setup with self-regulation and noisier light curves.

Table 5.6 presents the mean, minimum, and maximum values for F_1, ROC-OVO, and ROC-OVA across two sequence lengths (50 and 150). Our regularised approach demonstrates consistent improvements across most performance metrics. For *Data set I*, using two losses resulted in a noticeable improvement in the F_1 score, particularly for the longer sequence length of 150, where the mean F_1 increased from 55.4 (one loss) to 56.7 (two losses), with the minimum F_1 also improving from 49.4 to 50.3. Similarly, the ROC-OVO and ROC-OVA metrics improved under the two-loss configuration, particularly for

Table 5.4. **Classification performance metrics in testing sets for different loss functions across ten experiments.** 40,000 light curves were considered in the training set for each experiment. Red and black numbers indicate the worst and best performance for each metric, respectively.

Loss function	F_1			ROC-OVO			ROC-OVA		
	mean	min	max	mean	min	max	mean	min	max
Weighted cross entropy	**57.6**	**55.4**	**62.4**	88.5	86.8	89.7	87.5	85.7	89.0
Focal	56.7	52.4	60.0	88.5	**86.9**	89.6	87.4	**85.9**	88.5
Cross entropy	**55.0**	**48.9**	60.5	**88.7**	**85.5**	**91.0**	**87.7**	**84.0**	**90.1**

Table 5.5. **Performance statistics in testing sets for two different sn_ratio values.** The results include mean, minimum, and maximum values for F_1, ROC-OVO, and ROC-OVA. Ten experiments are summarised in each row, and 40,000 stars were used to train each model. Focal loss was used in these experiments. The row labelled "Two losses" indicates that the training was conducted using synthetic light curves. In contrast, the row labelled "One loss" represents results from training with the same representation and architecture but without using synthetic samples.

		sn_ratio	F_1			ROC-OVO			ROC-OVA		
			mean	min	max	mean	min	max	mean	min	max
Data set I	One loss	4	54.1	49.2	57.5	87.7	85.8	89.0	86.5	84.9	87.5
		6	55.2	51.0	58.8	88.0	86.6	89.2	86.7	84.8	88.0
	Two losses	4	57.6	54.9	61.2	88.5	87.2	89.5	87.5	86.2	88.5
		6	56.8	52.4	60.6	88.4	87.2	89.5	87.3	86.2	88.5
Data set II	One loss	4	59.1	50.2	62.4	89.4	87.0	90.8	90.2	88.3	91.4
		6	59.2	53.3	61.4	89.4	88.4	90.5	90.3	89.7	91.3
	Two losses	4	60.5	58.3	62.4	89.3	88.4	90.1	90.2	89.1	90.9
		6	59.3	55.3	63.6	89.6	88.4	91.1	90.6	89.6	91.8

the 150-length sequence, where ROC-OVO increased from 88.1 to 88.5, and ROC-OVA increased from 86.6 to 87.0. For *Data set II*, in addition to the self-regulated approach always outperforming one-loss, the performance delta is greater for light curves of length 50 compared to 150 (across all metrics). This means that there is more benefit in using regularisation with synthetic data when the sequences are shorter, which is what one would intuitively expect.

Table 5.6. **Performance metrics in testing sets for two different sequence lengths.** Ten experiments were conducted for each row using 40,000 stars for training.

		seq_length	F_1			ROC-OVO			ROC-OVA		
			mean	min	max	mean	min	max	mean	min	max
Data set I	One loss	50	53.1	49.3	57.1	85.0	83.6	86.7	83.6	82.7	85.3
		150	55.4	49.4	60.4	88.1	86.5	89.1	86.6	84.9	87.6
	Two losses	50	53.4	49.0	57.0	85.7	84.2	87.2	84.1	82.2	85.3
		150	56.7	50.3	60.1	88.5	87.4	89.8	87.0	85.7	88.3
Data set II	One loss	50	47.9	43.8	50.9	83.8	82.2	85.1	85.4	83.9	86.6
		150	55.2	51.1	59.3	87.6	86.0	89.0	88.6	87.3	89.8
	Two losses	50	50.4	40.8	56.5	84.3	78.1	85.8	85.8	79.8	87.0
		150	55.8	51.6	61.1	87.8	85.6	88.9	88.7	86.7	89.5

6. CHAPTER VI. CONCLUSIONS

6.1. Lessons learned

In this thesis, we have proposed three innovative methods to improve the reliability of classifiers under data shift conditions. These methods include one for assessing models and two for training models. Addressing data shift is crucial because it represents a significant challenge in real-world astronomy applications where the training data distribution differs from the testing data. Ensuring classifier reliability under these conditions is vital for maintaining accuracy and robustness in predictive models.

First, we introduced a novel method for assessing and ranking models by incorporating expert knowledge. This approach utilises informative priors based on deterministic physical rules, enabling the estimation of an informative marginal likelihood with well-known properties such as a model selector. This method eliminates the need for analytical proof of prior parameters, offering a straightforward and original alternative for incorporating prior knowledge in the model assessment of Bayesian classifiers without a time-consuming adaptation process. For evaluation purposes, we developed a method capable of introducing bias into a dataset according to the classification difficulty of each object. This procedure allowed us to test various strategies for model assessment under three bias conditions. Our results demonstrate that the informative marginal likelihood can identify more suitable models than non-informative cross-validated metrics.

Second, we proposed a novel approach to mitigating the data shift problem in tabular data by integrating symbolic knowledge into artificial neural networks (ANNs). The knowledge representation is based on characteristic ranges in astrophysically relevant parameters provided by domain experts. An ad-hoc training procedure was designed to inject this knowledge into ANNs, featuring a regularisation function, masks, and a two-phase back-propagation strategy. Experiments indicate that our method significantly improves performance across three complementary metrics (ACC, F1 score, and AUC) on the shifted testing set. When 1D signals were injected into the neural net, mean ACC on

the testing set improved by 2.0%. For 2D signals, a statistically significant improvement of 3.0% was observed in the AUC metric. Thus, our method offers a reliable alternative for classifying RR Lyrae stars, even in the presence of data shift.

Third, we introduced a novel approach to improve the reliability of variable star classifiers by addressing the challenges posed by data shift and class imbalance. Our methodology combines a self-regulated CNN with a PELS-VAE to dynamically generate synthetic light curves that address underlying data issues. This mechanism trains a classifier that can better generalise to unseen data. The self-regulated CNN architecture, a key component of our methodology, employs dual mask training to distinguish between real and synthetic data patterns effectively. The classifier learns from both sets without overfitting to the synthetic data, ensuring a balanced training process. The PELS-VAE model, conditioned on physical parameters, is crucial for generating realistic and highly informative synthetic light curves. These synthetic samples are derived from less probable physical parameter zones based on classification performance obtained from the confusion matrix in each epoch. Our experiments show that our approach improves classification reliability and enhances hyperparameter optimization by including performance on synthetic samples in the objective score. Additionally, our method demonstrates stable behaviour under different conditions, such as signal-to-noise ratio, loss functions, and sequence length. Our experiments also reveal that hyperparameter search can be improved using synthetic light curves.

6.2. Future work

To summarise, we have proposed three methods to mitigate the data shift problem, that operate by injecting knowledge during the training and model assessment. Each proposed method can be extended and applied in different scenarios. The data shift problem is a relevant drawback for data-based science in general since it is tough to detect and very complex to mitigate.

First, we propose an informative marginal likelihood outperforming traditional model selection approaches. It is interesting to study the current limit of marginal likelihood estimators; future work can also consider extensions such as (i) the use of other types of informative priors, e.g. priors over the proportion of classes or heavy-tailed distributions over BLR's weights; (ii) analysis of other time-series survey data sets; (iii) and the application of this approach to other classes (or subtypes) of variable stars.

The second proposal considers mask-based learning, which uses a mask of weights to learn expert knowledge. Future work includes the design and incorporation of new DRs focused on improving the classification of RR Lyrae subclasses in the presence of a data shift problem; for example, inject knowledge related to the light curve magnitude asymmetry from the available features (e.g., Rc, skew and gskew). Additionally, this procedure could be adapted to incorporate signals into a multiclass MLP classifier. Lastly, studying the convergence properties of this learning approach beyond empirical results would be interesting.

The last proposal offers a novel path to increase the reliability of variable star classifiers by leveraging recent advances in deep generative models. However, a few topics deserve discussion to extend these advances beyond this thesis.

Certain limitations in the current scenario constrain this approach's performance. Uncertainties in some physical parameters, such as metallicity, may hinder the PELS-VAE fitting process, thereby limiting the exploration of the physical space when generating new synthetic samples. Future technological and theoretical advances in physical parameter estimation should mitigate this issue. Additionally, this learning framework could be partially informed by spectroscopy-based estimations for some physical parameters according to dynamic PELS-VAE requirements.

We propose five policies to manage the interaction between the classifier and the generative model. These policies are based on greedy criteria, but other criteria could be proposed and tested. Future contributions can focus on searching for optimal policies for

each training set, and these policies can be learned during training. Moreover, these policies could be dynamic during the training phase. This additional learning layer can be very time-consuming, but it represents an interesting direction for further exploration.

Sampling methods can be improved by ensuring the generated samples explore the physical space better, focusing on specific low-represented zones beyond the proposed global density transformation. This exploration can also be linked with the prediction performance in different zones of physical parameters.

Extending the proposed methods to consider sub-types of variable stars can provide a more comprehensive classification framework. The models can achieve higher performance and reliability by refining the classification process to account for these sub-types. Designing a self-regulated classifier that considers a more extensive set of types and subtypes of stars is challenging as a research goal.

6.3. Final thoughts

In a data-driven environment where machine learning models assist in gaining new knowledge or support critical decisions, the data shift problem emerges as a significant topic. This issue can lead to incorrect, unfair, or biased decisions. Unfortunately, it is challenging because we cannot entirely comprehend the inner workings of a current machine learning model, nor can we precisely define when and how a dataset may be biased.

In astronomy, machine learning models play a crucial role in knowledge extraction and process automation. Given the nature of data collection in this field, with its inherent biases, implementing rigorous model validation before application is essential. Moreover, research focused on mitigating biases in the applications of machine learning models should be encouraged.

We have addressed the data shift problem on multiple levels. A common thread across these approaches is the integration of expert knowledge to improve optimization and statistical methods. We introduce innovative methods, including a novel marginal likelihood estimation procedure, a new mask-based neural network training, and a self-regulated training procedure. Each proposed method enhances classifier reliability.

The methods developed here have potential applications beyond variable star classification. As these models are refined, we anticipate their adaptability to other fields facing similar data-driven challenges. The convergence of knowledge-based, machine learning and generative model strategies establishes a promising foundation for future research, enhancing classifier reliability and performance under diverse and evolving data issues.

REFERENCES

Abdollahi, M., Torabi, N., Raeisi, S., and Rahvar, S. (2023). Hierarchical classification of variable stars using deep convolutional neural networks. *Iranian Journal of Astronomy and Astrophysics*, *2*.

Aguirre, C., Pichara, K., and Becker, I. (2019). Deep multi-survey classification of variable stars. *Monthly Notices of the Royal Astronomical Society*, *482*(4), 5078–5092.

Alcock, C., Allsman, R., Alves, D., Axelrod, T., Becker, A., Bennett, D., ... others (1997). The MACHO project large magellanic cloud microlensing results from the first two years and the nature of the galactic dark halo. *The Astrophysical Journal*, *486*(2), 697.

Arlot, S., Celisse, A., et al. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, *4*, 40–79.

Armstrong, D. J., Gómez Maqueo Chew, Y., Faedi, F., and Pollacco, D. (2013). A catalogue of temperatures for Kepler eclipsing binary stars. *Monthly Notices of the Royal Astronomical Society*, *437*(4), 3473–3481.

Atzmueller, M., and Sternberg, E. (2017). Mixed-initiative feature engineering using knowledge graphs. In *K-cap* (pp. 1–4).

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). DBPEDIA: a nucleus for a web of open data. In *The semantic web* (pp. 722–735). Springer.

Auvergne, M., Bodin, P., Boisnard, L., Buey, J.-T., Chaintreuil, S., Epstein, G., ... others (2009). The CoRoT satellite in flight: description and performance. *Astronomy & Astrophysics*, *506*(1), 411–424.

Baldassarre, L., Mourao-Miranda, J., and Pontil, M. (2012). Structured sparsity models for brain decoding from fMRI data. In *Prni* (pp. 5–8).

Barbary, K., Barclay, T., Biswas, R., Craig, M., Feindt, U., Friesen, B., ... Sofiatti, C. (2016). SNCosmo: Python library for supernova cosmology. *Astrophysics Source Code Library*, Astrophysics Source Code Library–1611.

Bassi, S., Sharma, K., and Gomekar, A. (2021). Classification of variable stars light curves

using long short term memory network. *Frontiers in Astronomy and Space Sciences*, *8*, 168.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., ... others (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Beaton, R. L., Freedman, W. L., Madore, B. F., Bono, G., Carlson, E. K., Clementini, G., ... others (2016). The Carnegie-Chicago Hubble program. i. An independent approach to the extragalactic distance scale using only population ii distance indicators. *The Astrophysical Journal*, *832*(2), 210.

Becker, I., Pichara, K., Catelan, M., Protopapas, P., Aguirre, C., and Nikzat, F. (2020). Scalable end-to-end recurrent neural network for variable star classification. *Monthly Notices of the Royal Astronomical Society*.

Benavente, P., Protopapas, P., and Pichara, K. (2017). Automatic survey-invariant classification of variable stars. *The Astrophysical Journal*, *845*(2), 147.

Biewald, L. (2020). *Experiment tracking with weights and biases.* Retrieved from https://www.wandb.com/ (Software available from wandb.com)

Blanton, M. R., and Roweis, S. (2007). K-corrections and filter transformations in the ultraviolet, optical, and near-infrared. *The Astronomical Journal*, *133*(2), 734.

Blei, D. M., and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, *112*(518), 859–877.

Bloemen, S., Groot, P., Woudt, P., Wolt, M. K., McBride, V., Nelemans, G., ... others (2016). MeerLICHT and BlackGEM: custom-built telescopes to detect faint optical transients. In *Ground-based and airborne telescopes vi* (Vol. 9906, p. 990664).

Bloom, J., Richards, J., Nugent, P., Quimby, R., Kasliwal, M., Starr, D., ... others (2012). Automating discovery and classification of transients and variable stars in the synoptic survey era. *Publications of the Astronomical Society of the Pacific*, *124*(921), 1175.

Bollacker, K., Cook, R., and Tufts, P. (2007). Freebase: A shared database of structured general human knowledge. In *Association for the advancement of artificial intelligence*

(Vol. 7, pp. 1962–1963).

Booth, R., and Jonas, J. (2012). An overview of the MeerKAT project. *African Skies*, *16*, 101.

Borghesi, A., Baldo, F., and Milano, M. (2020). Improving deep learning models via constraint-based domain knowledge: a brief survey. *arXiv preprint arXiv:2005.10691*.

Brown, A. G., Vallenari, A., Prusti, T., De Bruijne, J., Mignard, F., Drimmel, R., ... others (2016). Gaia Data Release 1-summary of the astrometric, photometric, and survey properties. *Astronomy & Astrophysics*, *595*, A2.

Budavári, T., Szalay, A., and Loredo, T. (2017). Faint object detection in multi-epoch observations via catalog data fusion. *The Astrophysical Journal*, *838*(1), 52.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. (2018). Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*.

Burhanudin, U., Maund, J., Killestein, T., Ackley, K., Dyer, M., Lyman, J., ... others (2021). Light-curve classification with recurrent neural networks for GOTO: dealing with imbalanced data. *Monthly Notices of the Royal Astronomical Society*, *505*(3), 4345–4361.

Cabrera, G. F., Miller, C. J., and Schneider, J. (2014). Systematic labeling bias: De-biasing where everyone is wrong. In *Icpr* (pp. 4417–4422).

Carrasco-Davis, R., Cabrera-Vives, G., Förster, F., Estévez, P. A., Huijse, P., Protopapas, P., ... Donoso, C. (2019). Deep learning for image sequence classification of astronomical events. *Publications of the Astronomical Society of the Pacific*, *131*(1004), 108006.

Castro, N., Protopapas, P., and Pichara, K. (2017). Uncertain classification of variable stars: Handling observational gaps and noise. *The Astronomical Journal*, *155*(1), 16.

Catelan, M., and Smith, H. (2015). *Pulsating stars*. Wiley-VCH, Weinheim.

Chambolle, A. (2004). An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, *20*(1), 89–97.

Christensen, N., and Meyer, R. (1998). Markov chain Monte Carlo methods for Bayesian gravitational radiation data analysis. *Physical Review D*, *58*(8), 082001.

Christensen, N., Meyer, R., Knox, L., and Luey, B. (2001). Bayesian methods for cosmological parameter estimation from cosmic microwave background measurements. *Classical and Quantum Gravity*, *18*(14), 2677.

Copperwheat, C., Marsh, T., Littlefair, S., Dhillon, V., Ramsay, G., Drake, A., ... others (2011). SDSS J0926+ 3624: the shortest period eclipsing binary star. *Monthly Notices of the Royal Astronomical Society*, *410*(2), 1113–1129.

Creevey, O. L., Sordo, R., Pailler, F., Frémat, Y., Heiter, U., Thévenin, F., ... others (2023). Gaia data release 3-astrophysical parameters inference system (apsis). I. Methods and content overview. *Astronomy & Astrophysics*, *674*, A26.

Dálya, G., Galgóczi, G., Dobos, L., Frei, Z., Heng, I. S., Macas, R., ... de Souza, R. S. (2018). GLADE: a galaxy catalogue for multimessenger searches in the advanced gravitational-wave detector era. *Monthly Notices of the Royal Astronomical Society*, *479*(2), 2374–2381.

Das, P., and Sanders, J. (2018). MADE: a spectroscopic mass, age, and distance estimator for red giant stars with bayesian machine learning. *Monthly Notices of the Royal Astronomical Society*.

Davis, J., and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *International conference on machine learning* (pp. 233–240).

Debosscher, J., Sarro, L., Aerts, C., Cuypers, J., Vandenbussche, B., Garrido, R., and Solano, E. (2007). Automated supervised classification of variable stars. *Astronomy & Astrophysics*, *475*(3), 1159–1183.

Debosscher, J., Sarro, L., López, M., Deleuil, M., Aerts, C., Auvergne, M., ... others (2009). Automated supervised classification of variable stars in the CoRoT programme-method and application to the first four exoplanet fields. *Astronomy & Astrophysics*, *506*(1), 519–534.

Dékány, I., and Grebel, E. K. (2020). Near-infrared search for fundamental-mode RR Lyrae stars toward the inner bulge by deep learning. *The Astrophysical Journal*, *898*(1), 46.

Deng, C., Ji, X., Rainey, C., Zhang, J., and Lu, W. (2020). Integrating machine learning

with human knowledge. *iScience*, 101656.

Ding, X., Song, Z., Wang, C., and Ji, K. (2024). Detection of contact binary candidates observed by TESS using the autoencoder neural network. *The Astronomical Journal*, *167*(5), 192.

Donoso-Oliva, C., Becker, I., Protopapas, P., Cabrera-Vives, G., Vishnu, M., and Vardhan, H. (2023). ASTROMER- A transformer-based embedding for the representation of light curves. *Astronomy & Astrophysics*, *670*, A54.

Drake, A., Djorgovski, S., Mahabal, A., Beshore, E., Larson, S., Graham, M., ... others (2009). First results from the Catalina real-time transient survey. *The Astrophysical Journal*, *696*(1), 870.

Efron, B., and Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, *92*(438), 548–560.

Ehrgott, M. (2005). *Multicriteria optimization* (Vol. 491). Springer Science & Business Media.

Elorrieta, F., Eyheramendy, S., Jordán, A., Dékány, I., Catelan, M., Angeloni, R., ... others (2016). A machine learned classifier for RR Lyrae in the VVV survey. *Astronomy & Astrophysics*, *595*, A82.

Eyer, L., Audard, M., Holl, B., Rimoldini, L., Carnerero, M., Clementini, G., ... others (2023). Gaia Data Release 3-Summary of the variability processing and analysis. *Astronomy & Astrophysics*, *674*, A13.

Eyer, L., and Mowlavi, N. (2008). Variable stars across the observational HR diagram. In *Journal of physics: Conference series* (Vol. 118, p. 012010).

Feast, M. (1996). The pulsation, temperatures and metallicities of mira and semiregular variables in different stellar systems. *Monthly Notices of the Royal Astronomical Society*, *278*(1), 11–21.

Ford, E., and Gregory, P. (2006). Bayesian model selection and extrasolar planet detection. *arXiv preprint astro-ph/0608328*.

Fortuin, V. (2022). Priors in Bayesian deep learning: A review. *International Statistical Review*, *90*(3), 563–591.

Gaia Collaboration, Eyer, L., Rimoldini, L., Audard, M., Anderson, R., Nienartowicz, K., ... others (2019). Gaia data release 2: Variable stars in the colour-absolute magnitude diagram. *Astronomy and Astrophysics*, *623*, A110.

García-Jara, G., Protopapas, P., and Estévez, P. A. (2022). Improving astronomical time-series classification via data augmentation with generative adversarial networks. *The Astrophysical Journal*, *935*(1), 23.

Gelman, A., Jakulin, A., Pittau, M., Su, Y., et al. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383.

Gelman, A., and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, *7*(4), 457–472.

Gelman, A., Simpson, D., and Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*(10), 555.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, *4*(1), 1–58.

Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*(1984), 20110553.

Gieren, W. P., Fouqué, P., and Gómez, M. (1998). Cepheid period-radius and period-luminosity relations and the distance to the Large Magellanic Cloud. *The Astrophysical Journal*, *496*(1), 17.

Glorot, X., and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249–256).

Golchi, S. (2019). Informative priors in Bayesian inference and computation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *12*(2), 45–55.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. The MIT Press, Cambridge, Massachusetts.

Gran, F., Minniti, D., Saito, R., Zoccali, M., Gonzalez, O., Navarrete, C., ... others (2016).

Mapping the outer bulge with RRab stars from the VVV survey. *Astronomy & Astrophysics*, *591*, A145.

Gregory, P., and Loredo, T. (1992). A new method for the detection of a periodic signal of unknown shape and period. *The Astrophysical Journal*, *398*, 146–168.

Griffiths, M. (2018). Explosive and eruptive variable stars. In *Observer's guide to variable stars* (pp. 155–174). Cham: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-030-00904-5_11 doi: 10.1007/978-3-030-00904-5_11

Groenewegen, M. (2020). Analysing the spectral energy distributions of galactic classical Cepheids. *Astronomy & Astrophysics*, *635*, A33.

Gronau, Q., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., ... Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of mathematical psychology*, *81*, 80–97.

Handler, G. (2009). Delta Scuti variables. In *Aip conference proceedings* (Vol. 1170, pp. 403–409).

Hanson, T. E., Branscum, A. J., Johnson, W. O., et al. (2014). Informative g-priors for logistic regression. *Bayesian Analysis*, *9*(3), 597–612.

Hartman, J. (2012). VARTOOLS: light curve analysis program. *Astrophysics Source Code Library*, Astrophysics Source Code Library–1208.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., ... Lerchner, A. (2017). β-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations (Poster)*, *3*.

Hoerl, A. E., and Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *42*(1), 80–86. Retrieved from http://www.jstor.org/stable/1271436

Hogg, D., and Foreman-Mackey, D. (2018). Data analysis recipes: Using Markov Chain Monte Carlo. *The Astrophysical Journal Supplement Series*, *236*(1), 11.

Ivanova, N. (2014). Binary evolution: Roche Lobe overflow and blue stragglers. In *Ecology of blue straggler stars* (pp. 179–202). Springer.

Jamal, S., and Bloom, J. S. (2020). On neural architectures for astronomical time-series classification with application to variable stars. *The Astrophysical Journal Supplement Series*, *250*(2), 30.

Jayasinghe, T., Kochanek, C., Stanek, K., Shappee, B., Holoien, T. W., Thompson, T. A., ... others (2021). The ASAS-SN catalogue of variable stars IX: The spectroscopic properties of galactic variable stars. *Monthly Notices of the Royal Astronomical Society*, *503*(1), 200–235.

Jayasinghe, T., Stanek, K., Kochanek, C., Shappee, B., Holoien, T. W., Thompson, T. A., ... others (2019). The ASAS-SN catalogue of variable stars–II. Uniform classification of 412 000 known variables. *Monthly Notices of the Royal Astronomical Society*, *486*(2), 1907–1943.

Jeffery, C., and Saio, H. (2016). Radial pulsation as a function of hydrogen abundance. *Monthly Notices of the Royal Astronomical Society*, *458*(2), 1352–1373.

Jenatton, R., Gramfort, A., Michel, V., Obozinski, G., Eger, E., Bach, F., and Thirion, B. (2012). Multiscale mining of fMRI data with hierarchical structured sparsity. *SIAM Journal on Imaging Sciences*, *5*(3), 835–856.

Jurcsik, J. (1995). Revision of the [Fe/H] scales used for globular clusters and rr lyrae variables. *Acta Astronomica*, *45*, 653–660.

Kafle, S., de Silva, N., and Dou, D. (2020). An overview of utilizing knowledge bases in neural networks for question answering. *Information Systems Frontiers*, *22*, 1095–1111.

Kaiser, N., Burgett, W., Chambers, K., and Denneau, L. (2010). Ground-based and airborne telescopes III. In *Society of photographic instrumentation engineers* (Vol. 7733, p. 77330E).

Kallrath, J., Milone, E. F., and Wilson, R. (2009). *Eclipsing binary stars: modeling and analysis* (Vol. 11). Springer.

Kang, Z., Zhang, Y., Zhang, J., Li, C., Kong, M., Zhao, Y., and Wu, X.-B. (2023). Periodic variable star classification with deep learning: Handling data imbalance in an ensemble augmentation way. *Publications of the Astronomical Society of the Pacific*, *135*(1051),

094501.

Kim, B., Ko, Y., and Seo, J. (2021). Novel regularization method for the class imbalance problem. *Expert Systems with Applications*, 115974.

Kim, D., and Bailer-Jones, C. (2016). A package for the automated classification of periodic variable stars. *Astronomy & Astrophysics*, *587*, A18.

Kingma, D., and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Kohavi, R., et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *14th international joint conference on artificial intelligence* (pp. 1137–1145).

Kolenberg, K., Fossati, L., Shulyak, D., Pikall, H., Barnes, T. G., Kochukhov, O., and Tsymbal, V. (2010). An in-depth spectroscopic analysis of the Blazhko star RR Lyrae * - i. characterisation of the star: abundance analysis and fundamental parameters. *Astronomy & Astrophysics*, *519*, A64. Retrieved from https://doi.org/10.1051/0004-6361/201014471 doi: 10.1051/0004-6361/201014471

Kovtyukh, V., Lemasle, B., Nardetto, N., Bono, G., da Silva, R., Matsunaga, N., ... Grebel, E. (2023). Effective temperatures of classical Cepheids from line-depth ratios in the H-band. *Monthly Notices of the Royal Astronomical Society*, *523*(4), 5047–5063.

Leimkuhler, B., Pouchon, T., Vlaar, T., and Storkey, A. (2020). Constraint-based regularization of neural networks. *arXiv preprint arXiv:2006.10114*.

Lendasse, A., Wertz, V., and Verleysen, M. (2003). Model selection with cross-validations and bootstraps—application to time series prediction with RBFN models. In *Artificial neural networks and neural information processing—icann/iconip 2003* (pp. 573–580). Springer.

Lin, T., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the ieee international conference on computer vision* (pp. 2980–2988).

Lucas, J., Tucker, G., Grosse, R., and Norouzi, M. (2019). Understanding posterior collapse in generative latent variable models. In *International conference on learning*

representations (pp. 1–16).

MacFarland, T., and Yates, J. (2016). *Mann–Whitney u test*. Springer.

MacKay, D. (1992). *Bayesian methods for adaptive models* (Unpublished doctoral dissertation). California Institute of Technology.

Mackenzie, C., Pichara, K., and Protopapas, P. (2016). Clustering-based feature learning on variable stars. *The Astrophysical Journal*, *820*(2), 138.

Mahabal, A., Sheth, K., Gieseke, F., Pai, A., Djorgovski, S., Drake, A., and Graham, M. (2017). Deep-learnt classification of light curves. In *Computational intelligence, ieee symposium series on* (pp. 1–8).

Marconi, M., Nordgren, T., Bono, G., Schnider, G., and Caputo, F. (2005). Predicted and empirical radii of RR Lyrae stars. *The Astrophysical Journal*, *623*(2), L133.

Martínez-Palomera, J., Bloom, J., and Abrahams, E. (2022). Deep generative modeling of periodic variable stars using physical parameters. *The Astronomical Journal*, *164*(6), 263.

Masci, F., Hoffman, D., Grillmair, C., and Cutri, R. (2014). Automated classification of periodic variable stars detected by the wide-field infrared survey explorer. *The Astronomical Journal*, *148*(1), 21.

Meng, X., and Wong, W. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 831–860.

Minniti, D., Lucas, P., Emerson, J., Saito, R., Hempel, M., Pietrukowicz, P., ... others (2010). VISTA variables in the via lactea (VVV): The public ESO near-ir variability survey of the milky way. *New Astronomy*, *15*(5), 433–443.

Murray, I., and Ghahramani, Z. (2005). *A note on the evidence and Bayesian Occam's razor* (Tech. Rep.). Gatsby Unit Technical Report.

Myung, I., and Pitt, M. (1997). Applying Occam's razor in modeling cognition: A bayesian approach. *Psychonomic bulletin and review*, *4*(1), 79–95.

Narayan, G., Zaidi, T., Soraisam, M., Wang, Z., Lochner, M., Matheson, T., ... others (2018). Machine-learning-based brokers for real-time classification of the LSST alert stream. *The Astrophysical Journal Supplement Series*, *236*(1), 9.

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

Naul, B., van der Walt, S., Crellin-Quick, A., Bloom, J., and Pérez, F. (2016). cesium: Open-source platform for time-series inference. *arXiv preprint arXiv:1609.04504*.

Neal, R. (2001). Annealed importance sampling. *Statistics and computing*, *11*(2), 125–139.

Neyshabur, B., Tomioka, R., and Srebro, N. (2014). In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*.

Nun, I., Pichara, K., Protopapas, P., and Kim, D.-W. (2014). Supervised detection of anomalous light curves in massive astronomical catalogs. *The Astrophysical Journal*, *793*(1), 23.

Nun, I., Protopapas, P., Sim, B., Zhu, M., Dave, R., Castro, N., and Pichara, K. (2015). FATS: feature analysis for time series. *arXiv preprint arXiv:1506.00010*.

Overstall, A., and Forster, J. (2010). Default Bayesian model determination methods for generalised linear mixed models. *Computational Statistics & Data Analysis*, *54*(12), 3269–3288.

Parviainen, H., Deeg, H., and Belmonte, J. (2013). Secondary eclipses in the CoRoT light curves-a homogeneous search based on Bayesian model selection. *Astronomy & Astrophysics*, *550*, A67.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825–2830.

Pérez-Galarce, F., Pichara, K., Huijse, P., Catelan, M., and Mery, D. (2021). Informative Bayesian model selection for RR Lyrae star classifiers. *Monthly Notices of the Royal Astronomical Society*, *503*(1), 484–497.

Pérez-Galarce, F., Pichara, K., Huijse, P., Catelan, M., and Mery, D. (2023). Informative regularization for a multi-layer perceptron RR Lyrae classifier under data shift. *Astronomy and Computing*, *43*, 100694.

Pichara, K., Protopapas, P., Kim, D., Marquette, J., and Tisserand, P. (2012). An improved

quasar detection method in EROS-2 and MACHO LMC data sets. *Monthly Notices of the Royal Astronomical Society*, *427*(2), 1284–1297.

Pichara, K., Protopapas, P., and León, D. (2016). Meta-classification for variable stars. *The Astrophysical Journal*, *819*(1), 18.

Pojmanski, G., and Maciejewski, G. (2005). The all sky automated survey. Catalog of variable stars. iv. 18ˆ h-24ˆ h quarter of the southern hemisphere. *Acta Astronomica, v. 55, pp. 97-122,(2005).*, *55*, 97–122.

Prša, A., and Zwitter, T. (2005). A computational guide to physics of eclipsing binaries. i. demonstrations and perspectives. *The Astrophysical Journal*, *628*(1), 426.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. (2009). *Dataset shift in machine learning*. The MIT Press, Cambridge, Massachusetts.

Raftery, A. E., Newton, M. A., Satagopan, J., and Krivitsky, P. (2006). *Estimating the integrated likelihood via posterior simulation using the harmonic mean identity.*

Raileanu, L. E., and Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, *41*(1), 77–93.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, *378*, 686–707.

Rao, R., Fung, G., and Rosales, R. (2008). On the dangers of cross-validation. an experimental evaluation. In *Proceedings of the 2008 siam international conference on data mining* (pp. 588–596).

Rasmussen, C., and Ghahramani, Z. (2001). Occam's razor. In *Advances in neural information processing systems* (pp. 294–300).

Richards, J. (2012). Overcoming sample selection bias in variable star classification. In *Astrostatistics and data mining* (pp. 213–221). Springer.

Richards, J., Starr, D., Brink, H., Miller, A., Bloom, J., Butler, N., ... Rice, J. (2011). Active learning to overcome sample selection bias: application to photometric variable

star classification. *The Astrophysical Journal*, *744*(2), 192.

Richards, R., and Groener, A. (2022). Conditional β-vae for de novo molecular generation. *arXiv preprint arXiv:2205.01592*.

Rubin, D. (1981). The Bayesian bootstrap. *The annals of statistics*, 130–134.

Ruffio, J., Mawet, D., Czekala, I., Macintosh, B., De Rosa, R., Ruane, G., ... others (2018). A Bayesian framework for exoplanet direct detection and non-detection. *The Astronomical Journal*, *156*(5), 196.

Saha, P., and Williams, T. (1994). Unfolding kinematics from galaxy spectra: A Bayesian method. *The Astronomical Journal*, *107*, 1295–1302.

Samus, N., Kazarovets, E., Durlevich, O., Kireeva, N., and Pastukhova, E. (2017). General catalogue of variable stars: Version GCVS 5.1. *Astronomy Reports*, *61*(1), 80–88.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Sesar, B., Hernitschek, N., Mitrović, S., Ivezić, H., Žand Rix, Cohen, J., Bernard, E., ... others (2017). Machine-learned identification of RR Lyrae stars from sparse, multi-band data: The ps1 sample. *The Astronomical Journal*, *153*(5), 204.

Sesar, B., Ivezić, Ž., Stuart, J. S., Morgan, D. M., Becker, A. C., Sharma, S., ... Oluseyi, H. (2013). Exploring the variable sky with LINEAR. II. halo structure and substructure traced by RR Lyrae stars to 30 kpc. *The Astronomical Journal*, *146*(2), 21.

Settles, B. (2009). *Active learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences.

Sharma, S. (2017). Markov Chain Monte Carlo methods for Bayesian data analysis in astronomy. *Annual Review of Astronomy and Astrophysics*, *55*, 213–259.

Shorten, C., and Khoshgoftaar, T. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, *6*(1), 1–48.

Smith, R. (2006). Cataclysmic variables. *Contemporary physics*, *47*(6), 363–386.

Sokolova, M., and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437.

Sooknunan, K., Lochner, M., Bassett, B. A., Peiris, H. V., Fender, R., Stewart, A., ...

Lahav, O. (2021). Classification of multiwavelength transients with machine learning. *Monthly Notices of the Royal Astronomical Society*, *502*(1), 206–224.

Spyroglou, I., Spöck, G., Chatzimichail, E., Rigas, A., and Paraskakis, E. (2018). A bayesian logistic regression approach in asthma persistence prediction. *Epidemiology, Biostatistics and Public Health*, *15*(1).

Sravan, N., Milisavljevic, D., Reynolds, J., Lentner, G., and Linvill, M. (2020). Real-time, value-driven data augmentation in the era of LSST. *The Astrophysical Journal*, *893*(2), 127.

Steeghs, D., Galloway, D., Ackley, K., Dyer, M., Lyman, J., Ulaczyk, K., ... others (2022). The gravitational-wave optical transient observer (GOTO): prototype performance and prospects for transient science. *Monthly Notices of the Royal Astronomical Society*, *511*(2), 2405–2422.

Suchanek, F., Kasneci, G., and Weikum, G. (2007). YAGO: a core of semantic knowledge. In *Proceedings of the 16th international conference on world wide web* (pp. 697–706).

Sugiyama, M., Krauledat, M., and MÃžller, K. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, *8*(May), 985–1005.

Svozil, D., Kvasnicka, V., and Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*, *39*(1), 43–62.

Szklenár, T., Bódi, A., Tarczay-Nehéz, D., Vida, K., Mező, G., and Szabó, R. (2022). Variable star classification with a multiple-input neural network. *The Astrophysical Journal*, *938*(1), 37.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, *58*(1), 267–288.

Trabucchi, M., Mowlavi, N., and Lebzelter, T. (2021). Semi-regular red giants as distance indicators-I. The period–luminosity relations of semi-regular variables revisited. *Astronomy & Astrophysics*, *656*, A66.

Trotta, R. (2008). Bayes in the sky: Bayesian inference and model selection in cosmology.

Contemporary Physics, *49*(2), 71–104.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525.

Tsang, B. T., and Schultz, W. (2019). Deep neural network classifier for variable stars with novelty detection capability. *The Astrophysical Journal Letters*, *877*(2), L14.

Udalski, A., Szymanski, M., Soszynski, I., and Poleski, R. (2008). The optical gravitational lensing experiment. final reductions of the OGLE-III data. *arXiv preprint arXiv:0807.3884*.

Uytterhoeven, K., Moya, A., Grigahcène, A., Guzik, J., Gutiérrez-Soto, J., Smalley, B., ... others (2011). The Kepler characterization of the variability among A-and F-type stars-I. General overview. *Astronomy & Astrophysics*, *534*, A125.

Valenzuela, L., and Pichara, K. (2017). Unsupervised classification of variable stars. *Monthly Notices of the Royal Astronomical Society*, *474*(3), 3259–3272.

VanderPlas, J. (2016). Gatspy: General tools for astronomical time series in Python. *Astrophysics Source Code Library*, Astrophysics Source Code Library–1610.

Van Laarhoven, P., and Aarts, E. (1987). Simulated annealing. In *Simulated annealing: Theory and applications* (pp. 7–15). Springer.

Vilalta, R., Gupta, K. D., and Macri, L. (2013). A machine learning approach to Cepheid variable star classification using data alignment and maximum likelihood. *Astronomy and Computing*, *2*, 46–53.

Von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., ... others (2021). Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, *35*(1), 614–633.

Wang, Y., Chen, M., and Kuo, P., Land Lewis. (2018). A new Monte Carlo method for estimating marginal likelihoods. *Bayesian analysis*, *13*(2), 311.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, *14*(Mar), 867–897.

Watanabe, T., Kessler, D., Scott, C., Angstadt, M., and Sripada, C. (2014). Disease prediction based on functional connectomes using a scalable and spatially-informed support vector machine. *Neuroimage*, *96*, 183–202.

Weinberg, M. (2013). Computational statistics using the Bayesian inference engine. *Monthly Notices of the Royal Astronomical Society*, *434*(2), 1736–1755.

Wright, E., Eisenhardt, P., Mainzer, A., Ressler, M., Cutri, R., Jarrett, T., ... others (2010). The wide-field infrared survey explorer (WISE): mission description and initial on-orbit performance. *The Astronomical Journal*, *140*(6), 1868.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. (2019). Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*.

Zhang, K., and Bloom, J. (2021). Classification of periodic variable stars with novel cyclic-permutation invariant neural networks. *Monthly Notices of the Royal Astronomical Society*, *505*(1), 515–522.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis*, *13*(2), 157–170.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, *67*(2), 301–320.

APPENDIXES

A. COMPLEMENTARY MATERIAL OF PUBLICATION 1

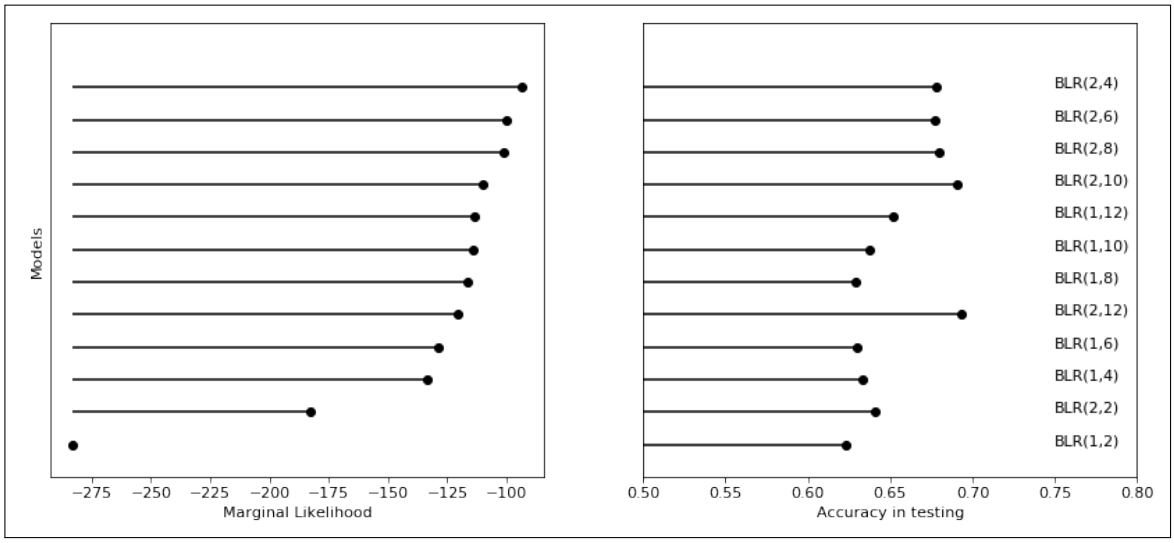


Figure A.1. **Comparison of model rankings on *rrlyrae-1* dataset sorted by the marginal likelihood BLR-IP(σ=10).**

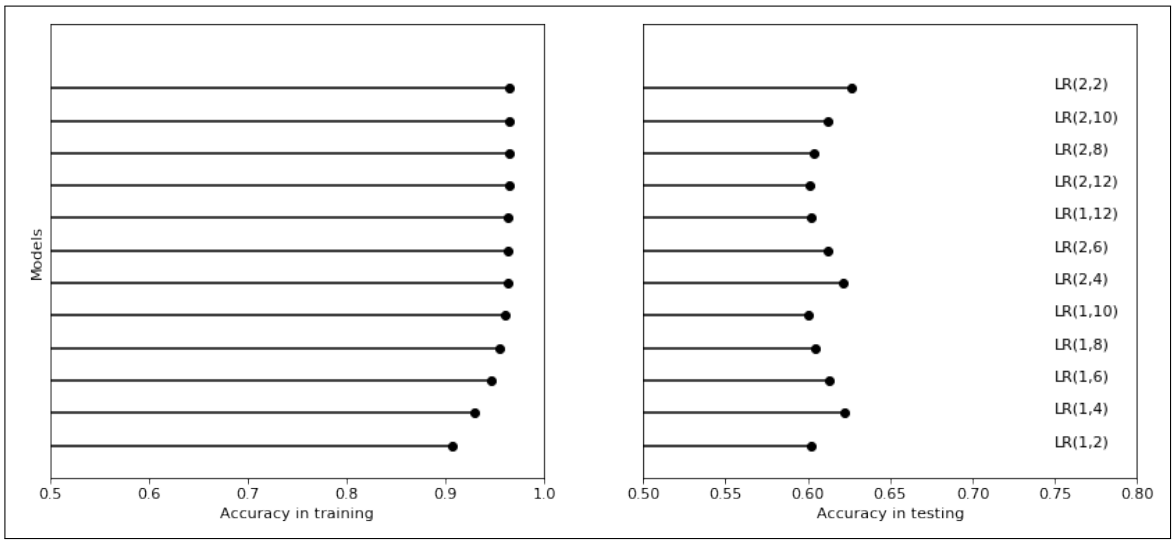


Figure A.2. **Comparison of model rankings on *rrlyrae-1* dataset sorted by a cross-validated (k=10) Accuracy for LR family of models.**

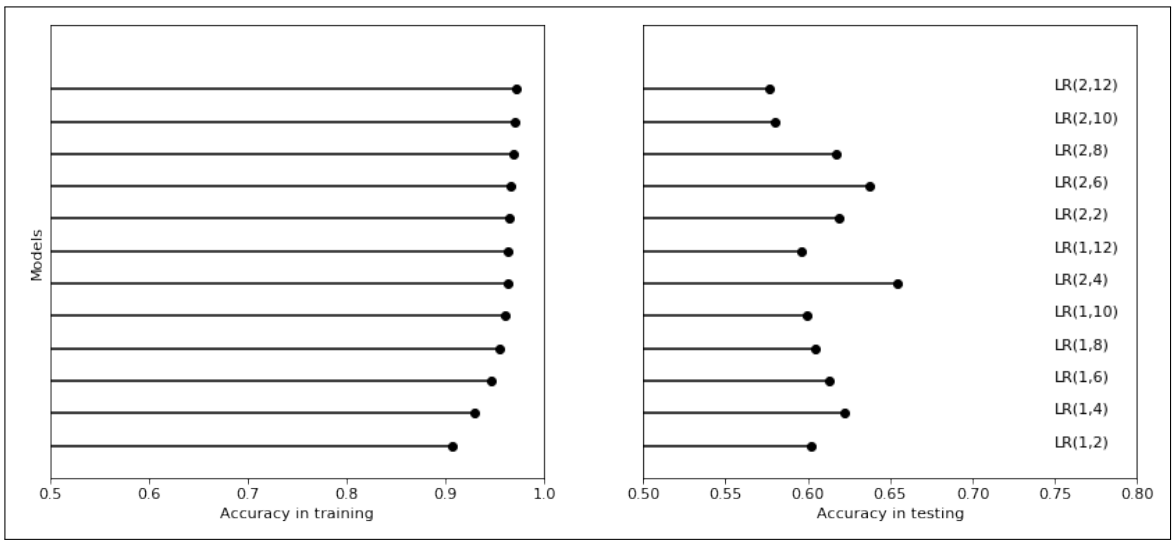


Figure A.3. **Comparison of model ranking on *rrlyrae-1* dataset sorted by a cross-validated (k=10) Accuracy for l_2-LR-100 family of models.**

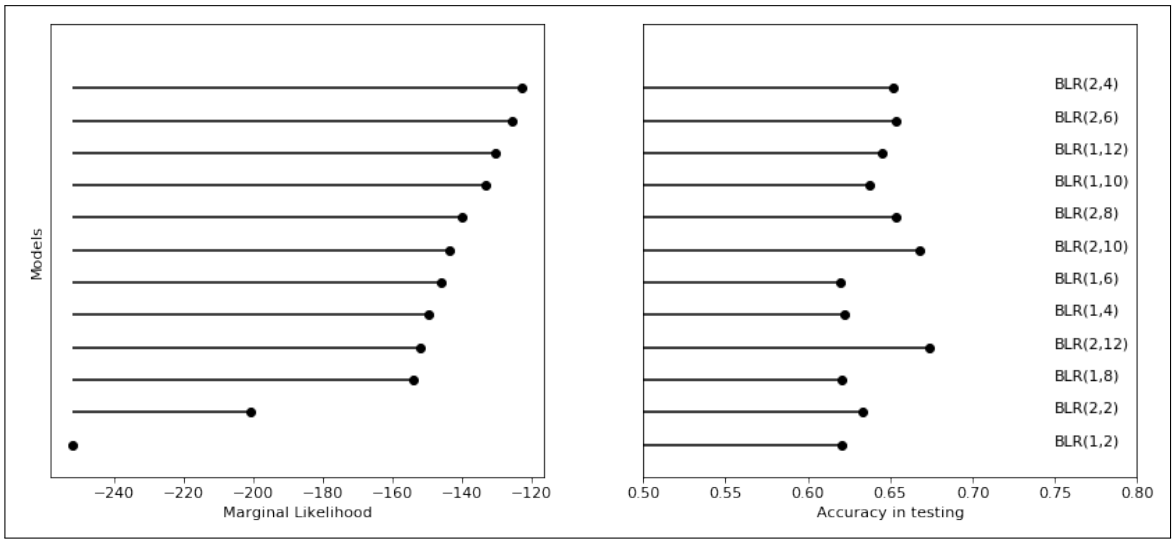


Figure A.4. **Comparison of model rankings on *rrlyrae-2* dataset sorted by the marginal likelihood BLR-IP(σ=10).**

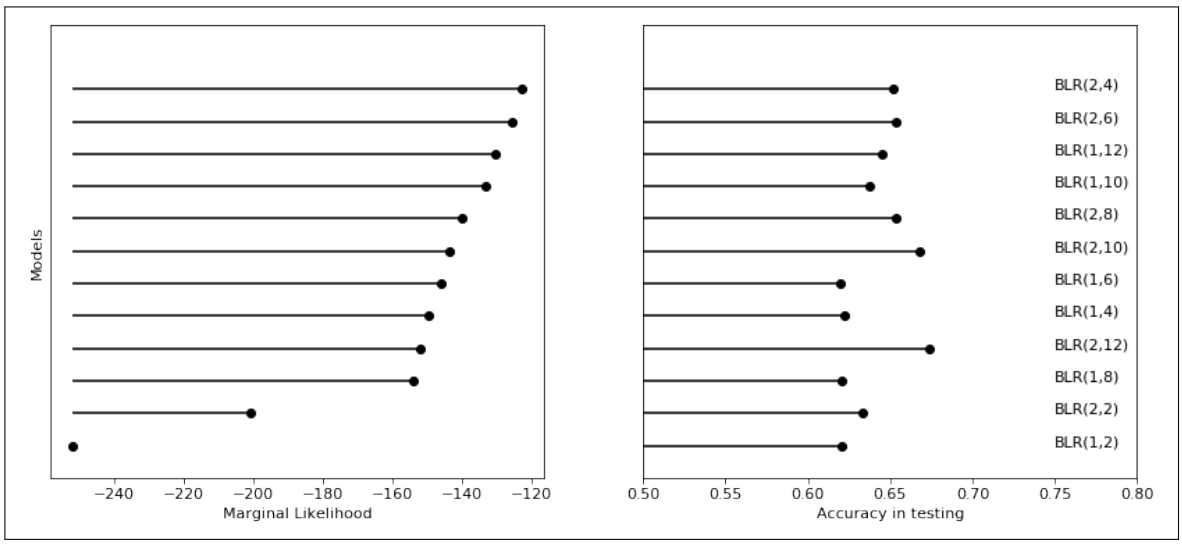
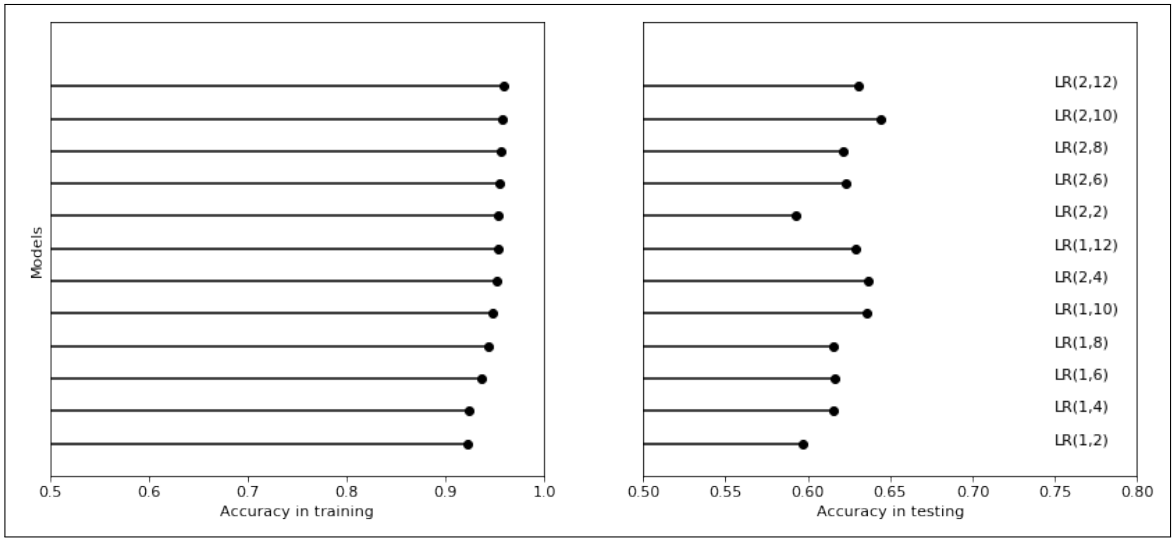


Figure A.5. **Comparison of model rankings on *rrlyrae-2* dataset sorted by a cross-validated (k=10) Accuracy for LR family of models.**



(a)

Figure A.6. **Comparison of model rankings on *rrlyrae-2* sorted by a cross-validated (k=10) Accuracy for l_2-LR-100 family of models.**

B. COMPLEMENTARY MATERIAL OF PUBLICATION 2

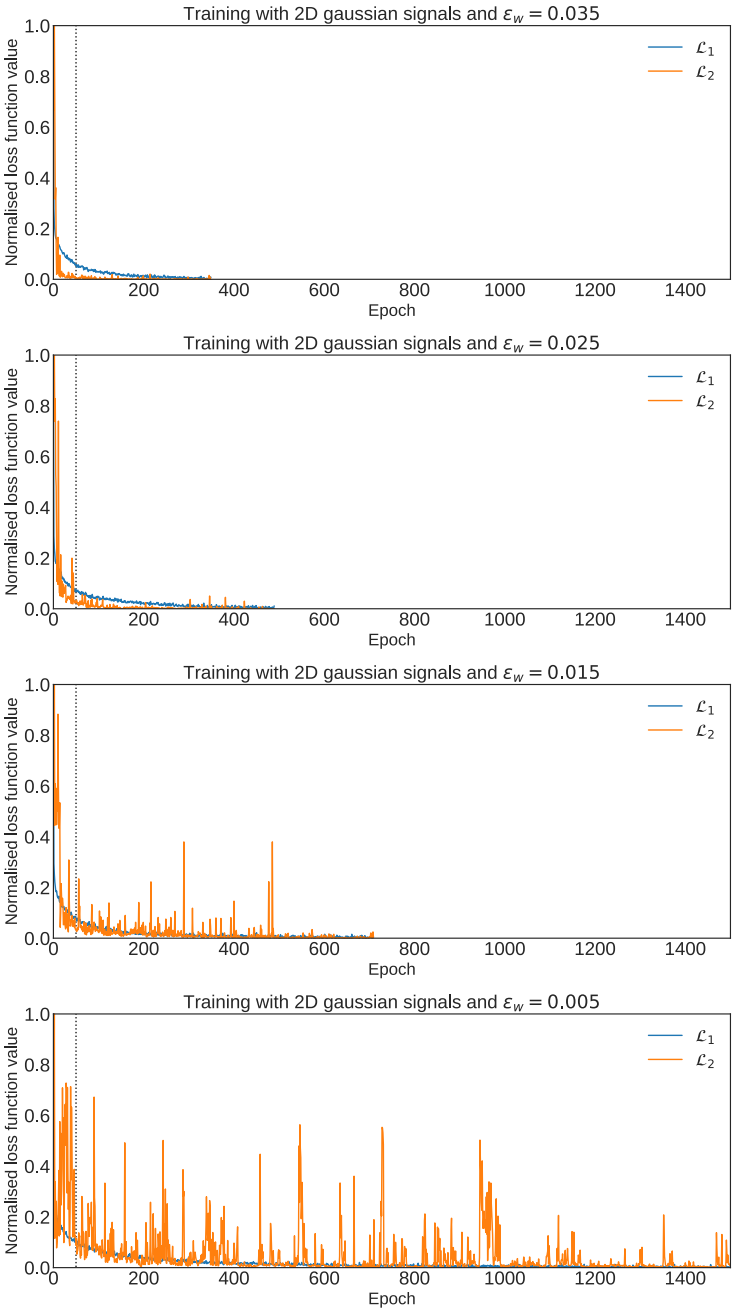


Figure B.1. **Convergence behaviour during the training process.** The dotted lines represent the end of the initial phase of the learning weight assignment. An early stopping criteria with patience equal to 10 is applied. Let $\mathcal{L}_1$ be the primary loss for classification and $\mathcal{L}_2$ represents the secondary loss for regularisation.

Table C.1. **Number N of missing astrophysical parameters.** Metallicity ([Fe/H]), effective temperature ($T_{\rm eff}$), period (P), absolute G magnitude (M_G), radius (R), and surface gravity ($\log g$).

Type	$N_{\rm [Fe/H]}$	$N_{T_{\rm eff}}$	N_P	N_{M_G}	N_R	$N_{\log g}$
CEP	1,735	339	0	0	1,734	1,735
DSCT	634	288	0	0	634	634
ECL	2,685	458	0	0	2,491	2,685
LPV	6,331	702	0	0	6,326	6,331
RRLYR	4,667	3,709	0	0	5,065	5,076

C. COMPLEMENTARY MATERIAL OF PUBLICATION 3

Table C.2. **Overview of approximate ranges of physical parameters for included variable stars in this study.**

Star Type	Complete Name	$T_{\rm eff}$ (K)	P (days)	M_G (G band)	$\log g$ (dex)	[Fe/H] (dex)	R ($R_\odot$)
RR Lyrae	RR Lyrae variables	6,200 - 6,800	0.30 - 0.90	0.30 to 1.00	2.5 - 3.5	-2 to 0	4.0 - 6.5
CEP	Classical Cepheids	5,000 - 8,000	1 - 200	-2 to +2	1.0 - 4.0	-0.5 to 0.5	5 - 50
DSCT	Delta Scuti variables	6,500 - 8,500	0.02 - 0.3	+0.5 to +4.0	3.5 - 4.5	-0.5 to 0.5	1.5 - 2.5
ECL	Eclipsing binaries	2,500 - 30,000	0.1 - 10,000	-3 to +10	Varies	Varies	Varies
LPV	Long-period variables	2,000 - 3,500	100 - 1,000	-1.5 to +2.0	0 - 2.0	-3 to 0	50 - 500

[1] References for physical parameters values (Feast, 1996; Marconi et al., 2005; Handler, 2009; Kallrath et al., 2009; Kolenberg, K. et al., 2010; Uytterhoeven et al., 2011; Copperwheat et al., 2011; Armstrong et al., 2013; Catelan and Smith, 2015; Groenewegen, 2020; Jayasinghe et al., 2021; Trabucchi et al., 2021; Kovtyukh et al., 2023; Eyer et al., 2023).

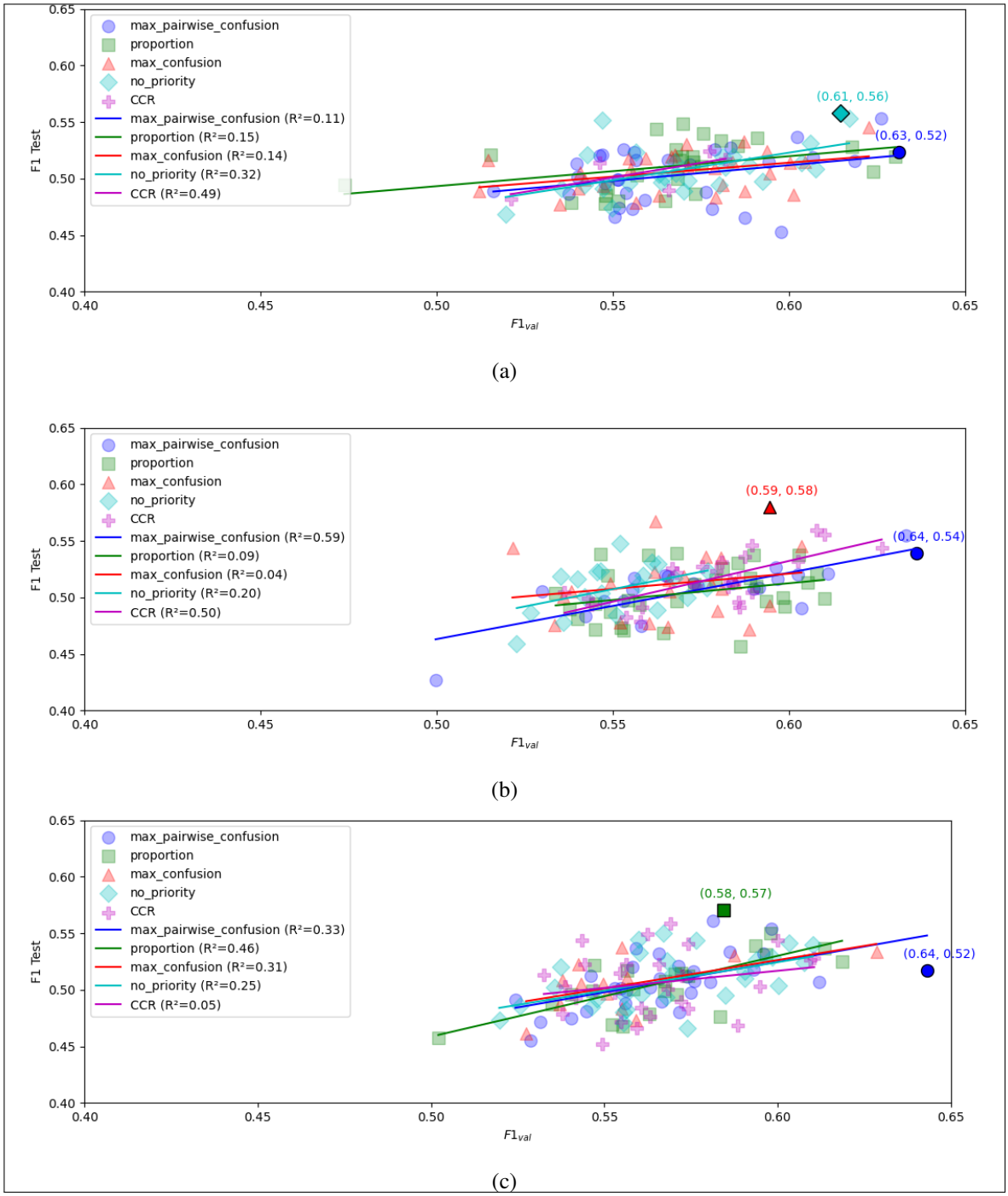


Figure C.1. **Hyperparameter search results using Bayesian optimisation with `Weights & Biases`.** Points represent different configurations, colour-coded by the policy employed for defining the number of samples for each class. Regression lines show relationships between the objective score (F_1^{val} or $F_1^{weighted}$) and F_1 test scores. 120 objective function evaluations were applied. (a) F_1^{val} is used as objective score. (b) $F_1^{weighted}$ is used as objective score, $F_1^{weighted} = 0.85{*}F_1^{val} + 0.15{*}F_1^{synthetic}$, where F_1^{val} is the macro weighted F_1 from the training set and $F_1^{synthetic}$ is the macro weighted F_1 from synthetic samples. (c) $F_1^{weighted}$ is used as objective score, $F_1^{weighted} = 0.7{*}F_1^{val} + 0.3{*}F_1^{synthetic}$.