Classification of Supernovae and Stars in the Era of Big Data and Artificial Intelligence

Jonathan Emrys Carrick

Physics

Department of Physics

Lancaster University

May 23, 2022

A thesis submitted to Lancaster University for the degree of Doctor of Philosophy in the Faculty of Science and Technology

Supervised by Prof. Isobel M. Hook

Abstract

In recent years, artificial intelligence (AI) has been applied in many fields of research. It is particularly well suited to astronomy, in which very large datasets from sky surveys cover a wide range of observations. The upcoming Legacy Survey of Space and Time (LSST) presents unprecedented big data challenges, requiring state-of-the-art methods to produce, process and analyse information. Observations of Type Ia supernovae help constrain cosmological parameters such as the dark energy equation of state, and AI will be instrumental in the next generation of cosmological measurements due to limited spectroscopic resources. AI also has the ability to improve our astrophysical understanding by perceiving patterns in data which may not be obvious to humans.

In this thesis we investigate how advanced AI methods can be used in classification tasks: to identify Type Ia supernovae for cosmology from photometry using supervised learning; by determining a low-dimensional representation of stellar spectra, and inferring astrophysical concepts through unsupervised learning.

In preparation for photometric classification of transients from LSST we run tests with different training samples. Using estimates of the depth to which the 4-metre Multi-Object Spectroscopic Telescope (4MOST) Time-Domain Extragalactic Survey (TiDES) can classify transients, we simulate a magnitude-limited training sample reaching $r_{\mathrm{AB}} \approx 22.5$ mag. We run our simulations with the software SNMACHINE, a photometric classification pipeline using machine learning. The machine-learning algorithms struggle to classify supernovae when the training sample is magnitude-limited as its features are not representative of the test

set. In contrast, representative training samples perform very well, particularly when redshift information is included. Classification performance noticeably improves when we combine the magnitude-limited training sample with a simulated realistic sample of faint, high-redshift supernovae observed from larger spectroscopic facilities; the algorithms' range of average area under ROC curve (AUC) scores over 10 runs increases from 0.547–0.628 to 0.946–0.969 and purity of the classified sample reaches 95% in all runs for 2 of the 4 algorithms. By creating new, artificial light curves using the augmentation software avocado, we achieve a purity in our classified sample of 95% in all 10 runs performed for all machine-learning algorithms considered. We also reach a highest average AUC score of 0.986 with the artificial neural network algorithm. Having real faint supernovae to complement our magnitude-limited sample is a crucial requirement in optimisation of a 4MOST spectroscopic sample. However, our results are a proof of concept that augmentation is also necessary to achieve the best classification results.

During our investigation into an optimised training sample, we assumed that every training object has the correct class label. Spectroscopy is a reliable method to confirm object classification and is used to define our training sample. However, it is not necessarily perfect and we therefore consider the impact of potential misclassifications of training objects. Taking the predicted error rates in spectroscopic classification from the literature, we apply contamination to a TiDES training sample using simulated LSST data. With the recurrent neural network from the software SuperNNova, we determine appropriate hyperparameters using a perfect, uncontaminated TiDES training sample and then train a model on its contaminated counterpart to study its effects on photometric classification. We find that a contaminated training sample produces very little difference in classification performance, even when increasing contamination to 5%. Contamination causes more objects of both Type Ia and non-Ia to be classified as Ia, increasing efficiency, but decreasing purity, with changes of less than 1% on

average. Similarly, we see a decrease of 0.1% in average accuracy, and no clear difference in AUC score, only varying at the fourth significant figure. These results are promising for photometric classification. Contaminated training appears to have little impact and propagation to cosmological measurements is expected to be minimal.

In a separate study, we apply deep learning to data in the European Southern Observatory (ESO) archive using an autoencoder neural network with the aim of improving similarity-based searches using the network's own interpretation of the data. We train the network to reconstruct stellar spectra by passing them through an information bottleneck, creating a low-dimensional representation of the data. We find that this representation includes several informative dimensions and, comparing to known astrophysical labels, see clear correlations for two key nodes; the network learns concepts of radial velocity and effective temperature, completely unsupervised. The interpretation of the other informative nodes appears ambiguous, leaving room for future investigation.

The results presented in this thesis emphasise the practical capabilities of AI in an astronomical context: Classification of astrophysical objects can be conducted through supervised learning using known labels, as well as unsupervised learning in a physics-agnostic process.

This thesis is dedicated to Skynyrd the cat.

Acknowledgements

I cannot begin to express my appreciation and thanks to my supervisor, Isobel Hook, an amazing scientist who has been helpful and understanding at every step of my PhD studies. Her knowledge and guidance has been invaluable in my research. I also sincerely thank her for organising my internship at ESO, which became an unforgettable experience.

Special thanks to the ESO amigos including Caz, Gabz, Connor, Vanessa and Anežka, to name a few - a fascinating, artistic and clever bunch of people who made my time at ESO some of the best months of my life.

Also, I'm extremely grateful to Nima Sedaghat, a friendly and approachable researcher who was more than happy to pass his deep knowledge of AI (no pun intended) down to me, not only while I worked with him at ESO, but beyond after my internship had finished as well.

I would also like to acknowledge and express sincere thanks to those individuals who helped me through tough times. I would not be the person I am now without those who spurred me on and helped provide me with confidence. In particular, I thank my friends: Robin Pawlett Howell for fun philosophical discussions and encouragement; David Thomas for general advice and assistance working with the Lancaster University High End Cluster; Lizi Swann for PhD advice and banter at research events; Angela Raj for being a great flatmate and offering mental support; Austin James for keeping me grounded in reality and always being a mate.

I am ever grateful to my family for general support and always being there whenever I needed a break from my studies.

I thank the ever-evolving Lancaster University astrophysics group for providing a friendly working environment, fun office times and pub trips.

Special thanks to the Lancaster University fencing society for providing me with new challenges and an enjoyable social life pre-pandemic.

Many thanks to fellow researchers from around the world who were friendly and willing to engage in this research, including members of TiDES and DESC, and notably: Anais Möller, for being dependable with discussions and software help; Emille Ishida for useful discussions, and organisation and support of research meetings; Kyle Boone for useful chats about software; Cat Alves for reliability, encouragement and chats about our PhD projects.

Thanks to all those who granted access to high-performance computing resources in order to carry out this research, including Lancaster University's High End Computing Cluster, the National Energy Research Scientific Computing Center and ESO's Rostam.

Many thanks also to STFC for providing the funding necessary to complete this PhD. The support included a Data Science studentship and multiple training workshops and events. My gratitude extends to the rest of the cohort of 4IR CDT students who made the experience very enjoyable.

Lastly, thanks to Lancaster University for the opportunity to pursue a PhD and for providing my student budget, enabling me to attend many research meetings, workshops and conferences, produce research posters, and also purchase equipment including the laptop that I have used throughout my PhD and written this thesis with.

Declaration

This thesis is my own work and no portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification at this or any other institute of learning.

Chapters 3, 4 and 5 include work published as a research paper in MNRAS (Carrick et al., 2021, MNRAS, 508, 1). This work covers optimising a TiDES training sample for photometric classification of supernovae. An overview of TiDES and its science goals is published in ESO's Messenger journal (Swann, Sullivan, Carrick et al. 2019, The Messenger, 175, 58) and relevant aspects are featured predominantly in Chapter 4. Chapter 7 covers the work undertaken during a data science placement at ESO, leading to a research paper published in MNRAS (Sedaghat, Romaniello, Carrick, Pineau, 2021, MNRAS, 501, 6026).

The work on photometric classification of supernovae in this thesis has been part of an official DESC Time Domain working group project: 'Optimising a spectroscopic training sample for photometric classification of transients'. Our publication was produced with contributions and approval from DESC.

Research and analyses done by others is given as follows:

- The machine learning software we use throughout our investigation of optimising a TiDES training sample is SNMACHINE (Lochner et al., 2016), created by Michelle Lochner and made available in DESC.

- The TiDES catalogue that we used to determine the magnitude limit based on 4MOST's capabilities (Chapter 4) was developed by Elizabeth Swann and Chris Frohmaier.
- We use the augmentation software AVOCADO (Boone, 2019) and adapt it to the dataset used in our analysis. This software was created by Kyle Boone who also helped us with our implementation.
- In our investigation of a contaminated training sample (Chapter 6), we use the classification framework SUPERNNOVA (Möller & de Boissière, 2020), created by Anais Möller.
- The simulated dataset of LSST transient light curves we used was created by others and provided by Maria Vincenzi.
- The autoencoder network in Chapter 7 was created by Nima Sedaghat (Sedaghat et al., 2021). Nima also developed the training method, including optimisation of λ for disentanglement and found the informative nodes for different latent space dimensionalities.
- Files containing physical labels for the HARPS objects were provided by François-Xavier Pineau.

"Compare yourself to who you were yesterday, not to who someone else is today." - Jordan Peterson

Contents

List of Figures

Relevant Publications by the Author

- "4MOST Consortium Survey 10: The Time-Domain Extragalactic Survey (TiDES)"; Swann, E.; Sulivan, M.; **Carrick, J.**; Hoenig, S.; Hook, I.; Kotak, R.; Maguire, K.; Nichol, R.; Smartt, S.; **2019, The Messenger, 175, 58.**
- "Machines learn to infer stellar parameters just by looking at a large number of spectra"; Sedaghat, N.; Romaniello, M.; **Carrick, J. E.**; Pineau, F.-X.; **2021, MNRAS, 501, 6026.**
- "Optimizing a magnitude-limited spectroscopic training sample for photometric classification of supernovae"; **Carrick, J. E.**; Hook, I. M.; Swann, E.; Boone, K.; Frohmaier, C.; Kim, A. G.; Sullivan, M.; (The LSST Dark Energy Science Collaboration); **2021, MNRAS, 508, 1.**

Chapter 1

Introduction

We are truly in the era of big data. Research institutes, companies and businesses are relying more and more on data-driven approaches to achieve their objectives. Thanks to advancements in computer hardware[1] and the world wide web it is easier than ever to access huge datasets and program scripts that can extract useful information – often described as *data mining*. This is particularly relevant to the astronomical community, as increasingly large sky surveys provide vast multi-dimensional observations of celestial objects (Baron, 2019; Boffin et al., 2019). Preparations for the Vera C. Rubin Observatory[2] (Ivezić et al., 2019) and its Legacy Survey of Space and Time (LSST) present perhaps the biggest of these challenges. The survey will accumulate ∼20 TB of data per night and, considering just one aspect of LSST science, an astounding 3–4 million supernova discoveries are anticipated over its 10-year duration. This will include hundreds of thousands of Type Ia supernovae. Compared to previous studies such as a Joint Light-curve Analysis of 740 spectroscopically confirmed Type Ia supernovae in Betoule et al. (2014) and a few thousand candidates from the recent Dark Energy Survey (Smith et al., 2020), it is clear that conducting Type Ia cosmology with LSST data will require state-of-the-art methods. Research in modern astrophysics is more commonly relying on artificial intelligence (AI) to generate and process

[1] The progress in: memory http://www.jcmit.net/memoryprice.htm; FLOPS https://ourworldindata.org/grapher/supercomputer-power-flops

[2] https://www.lsst.org/

whole datasets in order to classify, make predictions, and potentially discover new physics. This can immensely speed up the scientific process; 'big data is often available in real-time' (Mickaelian, 2020). In the pursuit of new astrophysics, development of AI methods should face the same scientific rigour, which itself helps pioneer advancements in big data analyses (Fluke & Jacobs, 2020).

1.1 Big data and artificial intelligence

One of the main benefits of using AI over traditional methods it that the algorithms can perceive patterns in data that humans cannot. Typical tasks in AI (with some overlap) are classification, regression, detection, image recognition and trend predictions. There are multiple approaches to achieve these. Broadly speaking, AI comprises machine learning and deep learning. Before we discuss the wider astrophysical context of this thesis, it is important to describe the background to the data science methods that are ubiquitous throughout this research.

1.1.1 Machine learning

Traditional data analysis methods are complemented incredibly well when incorporated into machine learning. Machine learning requires the programmer or user to specify what the machine is required to learn in order to process new information (Blum & Langley, 1997). Typically, these are defined as 'features' of the data; a model is trained to recognise features associated with data in a training set, and then apply what it has learnt about those features to new data in a test set. How the features are extracted from the data can also have a large effect on the outcome of the machine learning, which is seen when extracting features of supernova light curves in Lochner et al. (2016). See Li et al. (2017) for a comprehensive review on feature selection and class separation in many different applications. The basic principle of training and then testing data is prevalent across all forms of AI, but what distinguishes machine learning is the human input to tell the machine what form the useful information should be in.

Common machine learning models include, but are not limited to: support vector machines (Cortes & Vapnik, 1995), where a model tries to find the hyperplane in feature-space that best separates the data; decision trees, that can be incorporated into ensemble methods that combine weaker models into a strong one (Friedman, 2002); k-nearest neighbours (Altman, 1992), that works in a similar way to certain clustering algorithms by considering the distance between dataset objects defined in multi-dimensional feature-space and associating objects with their most similar neighbours; artificial neural networks, that are inspired by brain functionality where the strength of connections between neurons provides the required flow of information (Rosenblatt, 1960).

These algorithm examples are typically used with supervised learning, in which the training data is labelled (e.g. with a classification or category). The opposite is unsupervised learning (usually found in deep learning, clustering and dimensionality reduction tasks), in which training data is unlabelled and it is up to the algorithm to determine how the data is digested for its given task. In these cases, a model is developed from its own interpretation of training data and can then be applied to new data. Other options are semi-supervised learning, where only some of the data is labelled, and reinforcement learning, where the algorithm is rewarded or penalised for making the correct or wrong decisions (Géron, 2017).

Algorithms can be adjusted for the specific context and optimised by altering the hyperparameters that control how the information is handled. However, over-reliance on traditional machine learning models can introduce systematic biases, which is particularly problematic if there is uncertainty in the physics that is put into the machine.

1.1.2 Deep learning

It is impossible to eliminate all bias when creating a learning machine, however, arguably the best way to remove bias is to adopt a purely unsupervised deep learning approach where human input is minimal (Hinton & Sejnowski, 1999). Here we can apply the mantra of 'letting the data speak for itself'. In such models, the machine itself learns what features distinguish different types of data, and uses what it learns to complete its given task. Neural networks bridge the gap between

machine learning and deep learning. 'Deep' refers to the multiple layers within a network, although its main distinction from machine learning is *representation learning* (Bengio et al., 2013; Rumelhart et al., 1986; Sedaghat et al., 2021); in machine learning, the input is a set of features chosen by a human, whereas in deep learning, the network is only fed the original data and it will determine itself how to interpret the given information to construct its own representation.

Deep learning models consist of many various types of network, with architectures suited to different tasks. In convolutional neural networks information is processed via kernels with shared weights to take advantage of data that possesses a hierarchical structure or continuum (e.g. stellar spectra, Sedaghat et al. 2021), and is therefore generally suited for tasks such as image recognition (Simonyan & Zisserman, 2015). In recurrent neural networks, each layer corresponds to a point in a sequence, making these algorithms well-suited to temporal information (e.g. light curves, Möller & de Boissière 2020) and analysing real-time data such as audio (Sutskever et al., 2014). Autoencoders can combine different types of network. They compress data into a low-dimensional representation, and then attempts to reconstruct the original data in order to find a meaningful encoding at the information bottleneck (Tishby et al., 1999).

A common method to train deep neural networks is by minimisation of a loss function, where the amount of loss is calculated as the difference between the desired output and actual output, e.g. for a reconstruction task, the loss function is easily defined by calculating the difference between the original input and output. Minimisation of the loss function is achieved through gradient descent. Back propagation (Werbos, 1974) can be applied, calculating the weight-dependent gradient through each layer. Weights are updated in such a way that the high-dimensional loss function approaches a minimum, ideally the global minimum. Optimising training time and loss minimisation depends on the learning rate, which controls the change in weights at each iteration.

The similarities between how our minds work compared to machines is still not well established (Licata, 2015) and, whilst the successes of AI are incredibly impressive, AI is always prone to flaws as it is programmed by non-perfect humans[1].

[1]For now, although even the idea of AI creating new AI has to start with a human.

1.1.3 Human vs. artificial intelligence

Understanding of consciousness and how the mind works has become more important in recent decades with the rise of AI. In the nineteenth century, classical physics described a shared reality (before the revolutionary scientific advancements of relativity and quantum mechanics), yet the concept of the mind did not fit in this framework (Heisenberg, 1958). The ability of focused and sophisticated cognition appears to be a solely human trait, and a subjective experience (Tegmark, 2017). However, this breaks down if we consider machines to 'think' for themselves when they are programmed to learn. AI is an incredibly broad term, but this is particularly relevant when talking about neural network architectures, inspired by the network of biological neurons in the brain, i.e., the functionality of the mind.

To improve understanding of consciousness, there is a developing modern mathematical framework that describes consciousness based on information, defined as integrated information theory (Oizumi et al., 2014; Tononi, 2004) that can, in principle, be programmed into a machine. Computations within both man and machine seem to rely on how information is processed. Conversely, Roger Penrose famously said that consciousness is not a computation (Penrose, 1989), a rather controversial statement, particularly among computer scientists. Penrose's idea is that true conscious thought is non-deterministic due to quantum origins, and cannot be imitated by a machine. With respect to the mind, AI seems to most resemble unconscious thought, i.e., it learns things through computation and applies what it has learnt, but without consciously knowing it has done so[1].

We deal with both *deterministic* and *non-deterministic* algorithms (e.g. deterministic autoencoder vs. non-deterministic variational autoencoder), although this definition relies on the assumption that true randomness can be programmed[2]. Relaxing this assumption, even the deepest of artificial neural networks are deterministic, as a model given pre-determined information to learn from (with initial weights that can only ever be pseudo-random) will converge to the same state

[1] This topic probably requires its own thesis.

[2] "Any random number generated by a computer can be thought of as a member of a pseudo-random sequence. But in good implementations it usually can *draw* randomness from other things such as the clock, other external physical parameters, making it quite close to what we think is truly random." - Dr Nima Sedaghat.

(given a sufficient epoch of training), creating an unambiguous input-to-output function. We may not know the true nature of the human mind, but the programmable aspect of AI is fundamentally deterministic. This is not necessarily a bad thing; if an AI was developed that *was* inherently non-deterministic, then it would be impossible to say with absolute certainty how it learns anything. AI can be an incredibly useful tool, but its programmer must have a thorough understanding of how it reaches its conclusions.

It has been proven that machines have the potential to surpass human intelligence, with a notable turning point in the development of AI being the computer Deep Blue beating world chessmaster Garry Kasparov in 1997[1]. From then, it was clear that development of AI will lead to creation of machines that can see further and further beyond human comprehension. However, there is the caveat that, because they are originally programmed by humans, machines are prone to human error and artificial intelligence is still limited by human intelligence. Computers can be powerful, but one should be cautious; evidence of this is echoed by anybody who has experience in coding: "Why is it doing what I tell it, rather than what I want it to do?".

Common pitfalls when working with AI may be (but are certainly not limited to): optimising the wrong performance metrics (e.g. classification of a dataset with large class imbalance may have a high level of accuracy, but a low purity), where the correct choice of metrics should be dependent on the scientific context; not understanding whether an algorithm is actually learning from intrinsic information across the dataset, or is simply memorising specific instances in the training data (the question of 'if' a machine is learning something, rather than 'how'); not accounting for unknowns in the test set (a type of object may be present in the test set, but not in the training set); limited focus on homogenising the data (data preparation is arguably the most crucial aspect of a data-mining pipeline, e.g. consistent resolutions or boundary conditions), as the information supplied should be easily readable by the machine; over/under-fitting data, which occurs when a model's generalisation of training data does not reflect what we observe (often due to problems with the model, or with datasets that are either too small

[1] https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/

or 'noisy'). A lot of these instances are discussed throughout this thesis in their scientific contexts.

With all these potential pitfalls, it is paramount to have an understanding of what exactly an AI is learning and whether or not it is suited to its given task. AI is becoming the go-to tool for solving big data problems in the 21$^{\text{st}}$ century across all disciples in research and industry, but its users should absolutely have a firm grasp of how they are using it.

1.2 Background cosmology

Cosmology was one of the defining fields of modern physics in the 20$^{\text{th}}$ century. It developed from a philosophical question about the realm of our existence to a study of the whole Universe as a physical system. Nowadays we have astronomical observations that allow us to probe the biggest questions in physics. One cannot go through life without considering one's place in the Universe, and building on the pioneering physics of the last century will edge us ever closer to the truth of that place. The rest of this chapter introduces our current understanding of the Universe and the physical background motivating the work in this thesis.

Modern cosmology had its genesis when distant galaxies were first observed to all be moving away from us. The evidence of this is found in the characteristic absorption and emission lines of galaxy spectra. As galaxies recede, the lines are shifted towards redder wavelengths and the separation between them increased, reducing the frequency and energy of light waves, and creating a Doppler shift. We define this as redshift z, which is calculated as

$$z = \frac{\lambda_{observed} - \lambda_{emitted}}{\lambda_{emitted}}, \tag{1.1}$$

for observed and emitted wavelengths λ. The extent of this redshift provides a way of measuring the velocity of the galaxy, where redshift z and recession speed v are related by

$$1 + z = \sqrt{\frac{1 + v/c}{1 - v/c}}, \tag{1.2}$$

where c is the speed of light. This is often simplified to

$$z = \frac{v}{c}, \tag{1.3}$$

when special relativity can be ignored ($v \ll c$). This technique was first used to measure a galaxy's velocity in 1912 when Vesto Slipher found that the Andromeda galaxy (M31) is blueshifted and, hence, moving towards us (Slipher, 1912). It was not until 1929 that Edwin Hubble applied this technique to more distant galaxies that are not gravitationally bound to the Milky Way (Hubble, 1929). He found that the recession speed of galaxies is proportional to their distance away d, something we refer to now as the *Hubble-Lemaître law*:

$$v = H_0 d, \tag{1.4}$$

where H_0 is the Hubble constant[1]. Not only are galaxies moving away from us, but also from each other. To emphasise this point, we often write the Hubble-Lemaître law in a more general vector form:

$$\mathbf{v} = H_0 \mathbf{r}, \tag{1.5}$$

which holds true between comparison of galaxies with positions $\mathbf{r}$ and velocities $\mathbf{v}$; for an observer anywhere in the Universe, the same phenomenon would be seen as there is no special position or direction, i.e. there is no 'centre' to the Universe. This is succinctly described by the *cosmological principle*, which states that on large enough scales the Universe is homogenous and isotropic. When we describe galaxies as 'moving away', it is not that they are travelling through space, it is that the space between galaxies is expanding, and we can translate this to observed expansion everywhere in the Universe. It is then not a vast leap in logic to conclude that there must have been a moment from which expansion began – this is what we refer to as the Big Bang. An attempt to understand the nature of expansion is what constitutes cosmology. Studying distant light sources and their motion offers a glimpse into the mechanisms of the entire Universe.

[1] H_0 is not actually a constant, but the present-day value of the time-dependent Hubble parameter: $H_0 = H(t = t_0)$

1.2.1 The ΛCDM model

For a distant object with a certain flux f, its luminosity distance d_L is the distance that it appears to have based on its luminosity L. This assumes a reduction of light intensity following the inverse square law (Liddle, 2003), where observed flux is given by

$$f = \frac{L}{4\pi d_L^2}. \tag{1.6}$$

The luminosity distance is different to the actual *physical* distance due to expansion and any deviations from a perfectly flat geometry. Expansion causes photons to lose energy and rate of emission to drop, resulting in objects appearing dimmer and seeming further away. Assuming a flat geometry across the Universe, the physical distance is given by

$$d = \frac{d_L}{(1+z)}. \tag{1.7}$$

As Type Ia supernovae have a known luminosity (discussed in § 1.2.3.1), a luminosity distance can be calculated from their flux. In the late 1990s, two teams, the High-z Supernova Search Team and the Supernova Cosmology Project, independently found that at given redshifts, observed Type Ia supernovae appeared fainter than expected, implying that the expansion rate of the Universe is actually accelerating (Perlmutter et al., 1999; Riess et al., 1998). To explain the observed relationship between supernova luminosity distances and redshift, cosmological models required that, in a geometrically flat universe (where the total density parameter is $\Omega_{tot} = 1$), the density parameter[1] of matter is $\Omega_m \approx 0.3$.

The discrepancy could not be explained without cosmological models including an additional energy density term Knop et al. (2003). Assuming that this contribution has a constant energy density, it is referred to as the *cosmological constant*, represented by Λ. This lead to the ΛCDM model of the Universe, in

[1]The density parameter is defined as a fraction of the critical density ($\rho_c \equiv 3H_0^2/8\pi G$), i.e. the density of a perfectly flat Universe: $\Omega_i \equiv \rho_i/\rho_c$. Hence, in a geometrically flat universe, $\Omega_{tot} = 1$, as opposed to spherical $\Omega_{tot} > 1$ or hyperbolic $\Omega_{tot} < 1$ geometries.

which the total energy density is dominated by two terms – the energy densities of a cosmological constant and matter:

$$\rho_{tot} = \rho_\Lambda + \rho_m. \tag{1.8}$$

The matter (m) term consists of primarily cold dark matter (CDM), as opposed to baryonic matter that makes up observables such as stars, planets and nebulae. Cold dark matter is named as such as it is weakly interacting; its existence can only be inferred by gravitational effects and is not seen through direct interaction with ordinary matter or light. Dark matter has been postulated to explain galactic motion in clusters (Zwicky, 1937) and also the rotation curves of galaxies (Rubin & Ford, 1970), to account for 'missing' mass following the established knowledge of gravity. Observations also suggest that dark matter has played a crucial role in the growth of large scale structure (galaxies and galaxy clusters): structure formation relies on gravitational attraction in the potential wells caused by primordial density perturbations, and observed structure can only be explained by the existence of dark matter (Einasto, 2001).

The cosmological constant (Λ) energy density term corresponds to a constant, non-zero vacuum density that is defined by

$$\rho_\Lambda \equiv \frac{\Lambda c^2}{8\pi G} \tag{1.9}$$

where G is the Newtonian gravitational constant.

The luminosity distance depends on cosmological parameters as follows:

$$d_L(z) = \frac{c(1+z)}{H_0} \int_0^z \frac{dz'}{E(z')}, \tag{1.10}$$

where

$$E(z') = \sqrt{\Omega_m (1+z')^3 + \Omega_\Lambda}. \tag{1.11}$$

Under the assumption of a flat geometry in the ΛCDM model, Ω_Λ can be replaced by $1 - \Omega_m$, allowing Ω_m to be numerically evaluated.

Eq. 1.10 can be expanded in a Taylor series to give

$$d_L(z) = \frac{c}{H_0}\left[z + \frac{1}{2}(1 - q_0)z^2 + ...\right] \tag{1.12}$$

where q_0 is the deceleration parameter. Luminosity distance measurements of Type Ia supernovae imply that $q_0 < 0$, providing evidence that expansion is accelerating. q_0 can be expressed in terms of the different forms of energy density:

$$q_0 = \frac{1}{2}\sum_i (1 + 3w_i)\Omega_i, \tag{1.13}$$

where w_i is the equation of state of energy component i. The equation of state describes how the pressure p_i of a material in our Universe depends on its energy density, given by

$$w_i \equiv \frac{p_i}{\rho_i c^2}. \tag{1.14}$$

For the constant Λ contribution, its equation of state is $w_\Lambda = -1$. The speed of light c is a positive constant, and energy density ρ_Λ cannot be negative, hence it is its negative pressure that appears to counteract gravity and drive an accelerated expansion. In this model, matter is a pressureless fluid, meaning that $w_m = 0$. Hence, current estimates suggest that $q_0 \approx -0.55$ (Camarena & Marra, 2020).

The ΛCDM model relies on the assumption that Albert Einstein's general theory of relativity is correct, as it is derived using the Friedmann-Lemaître-Robertson-Walker metric. Close to home we have strong observational evidence that this is the case, e.g., the precession of Mercury's perihelion, the bending of starlight around the Sun (Dyson et al., 1920)[1], and also on satellites: atomic clocks need to take into account the effects of relativity for accurate global positioning systems (GPS). In recent years we have also obtained evidence from much further away, such as the first detection of gravitational waves (Abbott et al., 2016), and the first image of a black hole (Event Horizon Telescope Collaboration et al., 2019), confirming predictions from general relativity. However, extending relativity across the whole Universe requires energy beyond that of 'ordinary' baryonic matter.

[1]This was first observed by Arthur Eddington during the solar eclipse of 1919, launching Einstein's theory into the public eye.

If we calculate the theoretical age of the Universe assuming that there exists no Λ contribution, and using the H_0 value from Planck Collaboration et al. (2020), we compute an age of $t_0 = 9.64 \pm 0.06$ Gyr. We know this cannot be correct as the Universe cannot be younger than any of the physical objects it contains. Perhaps most convincingly, globular clusters were some of the first structures to form in the Milky Way; Hansen et al. (2002) used the white dwarf cooling sequence to constrain the age of globular cluster M4 to be 12.7 ± 0.7 Gyr. Combined with observational evidence of accelerating expansion, there is clearly a significant gap in our understanding of the Universe and, using the well-established basis of general relativity, one must conclude that there is an energy contribution other than matter influencing expansion. This form of energy does not appear to emit any detectable radiation, nor can it be inferred through gravitational effects. It is not currently known whether this energy density is exactly constant in time. To describe this energy generally (as not necessarily constant), it is often defined as *dark energy* (Huterer & Turner, 1999) due to its elusive nature.

1.2.2 Dark Energy

Dark energy appeared to have overtaken matter as the most-dominant contribution relatively recently ($z \lesssim 0.5$; Huterer & Shafer 2017), causing late-time cosmic acceleration. Depending on the assumed model, dark energy's effects on cosmic dynamics is different at high redshifts, although has been a significant contribution ($\geq$5% of the total energy density) out to $z \approx 2.5$ (Linder, 2021). One of the main issues surrounding dark energy as a cosmological constant is that its appearance as a vacuum energy density (Eq. 1.9), with $\rho_\Lambda \simeq 10^{-47}\text{GeV}^4$, is wildly incompatible with the standard model of particle physics, that predicts a density of $\rho_{vac} \simeq 10^{74}\text{GeV}^4$. To solve this massive discrepancy from a cosmological perspective, we can consider the cosmological dynamics that are dictated by the Einstein equations:

$$G_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}, \tag{1.15}$$

where either gravity needs modifying in a departure from general relativity (modifying the left-hand side – the Einstein tensor $G_{\mu\nu}$), or we need to consider alternative dark energy models (specific forms of the energy-momentum tensor $T_{\mu\nu}$ with a negative pressure on the right-hand side, Amendola & Tsujikawa 2010), such as the so-called *quintessence* (dynamical dark energy; Caldwell et al. 1998; Ratra & Peebles 1988).

If we relax the assumption that dark energy has a constant density contribution across cosmic time, its equation of state can be expressed as

$$w_{\rm DE}(z) = w_0 + w_1 \frac{z}{(1+z)^p}, \tag{1.16}$$

following the suggested parametrisation in Linder (2003b) where $p = 1$ and generalised in Jassal et al. (2005) to also consider $p = 2$. This variable form of the dark energy equation of state (Eq. 1.16, with $p = 1$) is also often expressed in terms of the scale factor a, a relative measure of the size of the Universe:

$$w(a) = w_0 + w_a(1 - a). \tag{1.17}$$

a is related to z following $a_0/a = 1 + z$ and at the present $a_0 = 1$. This means that, e.g., at $z = 2$ the Universe was 1/3 of its current size. Observing an object at redshift z is seeing the Universe when it was $1/(1 + z)$ of its present size and in the time the object's light has taken to reach us, it has been redshifted by a factor of $1 + z$ (Liddle, 2003).

The Friedmann continuity equation is a differential equation for energy density as followed:

$$\dot{\rho} + 3H\left(\rho + \frac{p}{c^2}\right) = 0. \tag{1.18}$$

We hence find that the energy density of a contribution i depends on its equation of state w_i (Eq. 1.14) and evolves with expansion following

$$\rho_i \propto a^{-3(1+w_i)}. \tag{1.19}$$

For matter, $w_m = 0$ and so its energy density reduces with expansion following

$\rho_m \propto a^{-3}$. The nature of the dark energy equation of state will tell us how it changes with respect to expansion. For $w_i = -1$, ρ_i is described by a cosmological constant (in which case we would use $i = \Lambda$).

Solving the continuity equation (Eq. 1.18) with the dark energy equation of state in the form given in Eq. 1.16 and $p = 1$ gives the more generalised

$$E(z') = \sqrt{\Omega_m(1+z')^3 + \Omega_{\rm DE}(1+z')^{3(1+w_0+w_1)}\exp\left(-3w_1\frac{z'}{1+z'}\right)} \tag{1.20}$$

which can go into calculation of the luminosity distance (Eq. 1.10) and can therefore be used to test the validity of a variable dark energy.

1.2.2.1 Cosmic inflation

It is believed that in the very early Universe cosmic inflation occurred, an epoch of extreme accelerated expansion (Guth, 1981). This was likely driven by a form of dark energy (Liddle & Lyth, 2000), offering further evidence to the presence of an unknown energy contribution present in the Universe.

The first evidence is seen in the cosmic microwave background radiation (CMB) that permeates the Universe, as regions thought to be causally disconnected are in thermal equilibrium (Fixsen et al., 1996). Inflation is a mechanism in which these regions could grow beyond the limit of causal contact. This does not violate relativity, as the speed of light is a limit to transport of matter and energy, but does not apply to space itself and expansion of space is not bound by this limit.

Inflation also offers an explanation of the observed flat geometry. Flat geometry appears to be an unstable solution, in which any slight deviation from perfectly flat geometry will increase rapidly in time. However, in a period of cosmic inflation, any deviation from flatness will be inflated away, and the geometry will resemble flatness on the scale of the observable Universe.

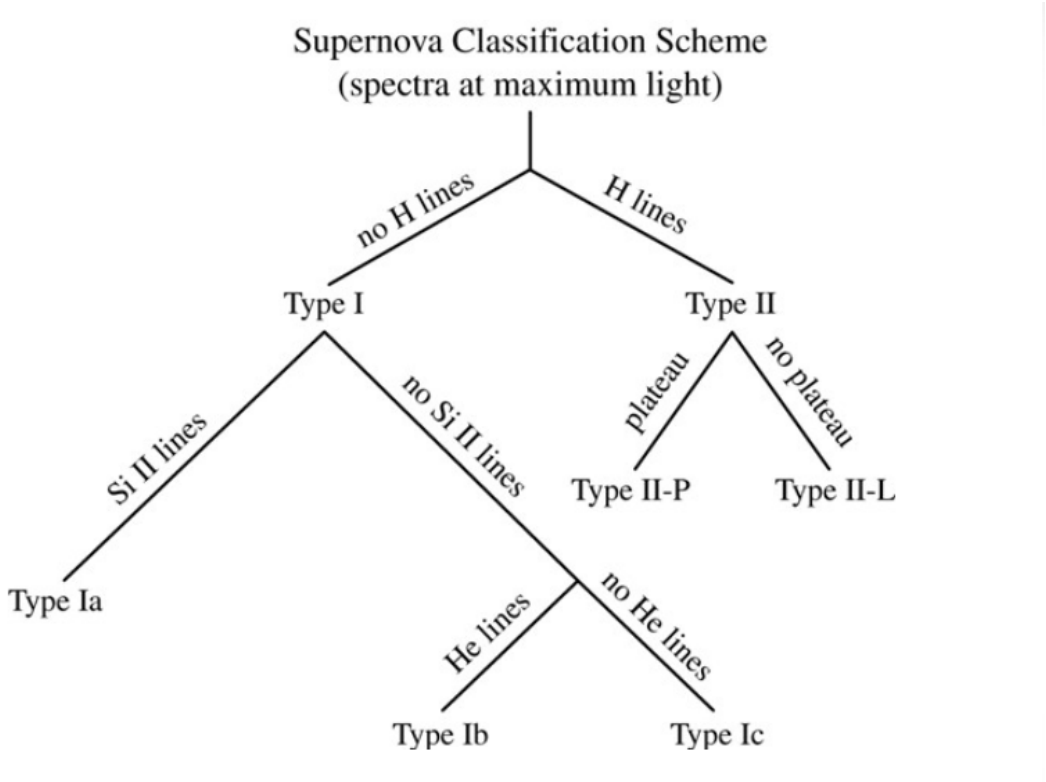


Figure 1.1: A basic summary of supernova classification, based on spectral signatures in the optical regime at maximum light and light curves for Type II. Credit: Carroll & Ostlie (2007).

1.2.3 Supernovae types

Supernovae are an incredibly diverse group of celestial objects, varying across progenitor models, explosion mechanics and chemical signatures (Filippenko, 1997). Such events have the ability to shine as bright as their host galaxies. They can be generally be split into two distinct groups: Type I and Type II.

Type I supernovae do not exhibit any hydrogen lines in their spectra, indicating that the stars have been stripped of their hydrogen envelopes, whereas spectra of Type II supernovae have strong hydrogen lines. These can be split further into subtypes. The presence of silicon is associated with Type Ia supernovae. Type I supernovae with no silicon are either Ib if they have helium or Ic otherwise. Type II supernovae are split based on their light curve profiles: Type II-P (plateau) or Type II-L (linear). Doggett & Branch (1985) is a comparative study of light

curves of different supernovae, presenting typical examples of the above types. These supernova types are summarised in a decision tree in Fig. 1.1 taken from Carroll & Ostlie (2007).

Collectively, Type Ib, Type Ic and Type II supernovae are known as core-collapse supernovae. At the end of a massive ($\gtrsim 8M_\odot$) star's life, the combination of photodisintegration of iron and electron capture quickly removes the stellar core's support, resulting in a rapid collapse. The excessive core density causes the strong nuclear force to enter its repulsive regime, sending a rebounding shockwave outwards and ejecting the star's envelope. This process is well-simulated, although occurrence and nature of a core-collapse explosion seems to be sensitive to stellar masses and metallicities (Heger et al., 2003). This does not cover all potential core-collapse supernovae, as there are a multitude of peculiar and separate subclasses, e.g. Type IIn which possesses narrow emission lines (Filippenko, 1997).

Type IIn supernovae originate from explosion ejecta interacting with dense circumstellar material. Aside from broader hydrogen lines characteristic of Type II supernovae, the presence of narrow emission lines indicates emission from this circumstellar material, as the radiative shock of the ejecta passes through the surrounding gas (Salamanca et al., 2002). These include lines with a P-Cygni profile, where there is absorption and emission of the same spectral line in the gas expanding away from the star (e.g. stellar winds, Scuderi et al. 1994).

Type Ia supernovae are a result of thermonuclear explosion and constitute a relatively homogenous subclass, making them the most useful supernovae for studying cosmic expansion (§ 1.2.3.1). However, there exists some ambiguities in classification due to variations within this subclass. Cosmologists use 'normal' Type Ia supernovae. Other Type Ia supernovae exist such as the bright 91T-like and faint 91bg-like and super-Chandrasekhar transients (Taubenberger, 2017). While relatively uncommon, their presence is an obstacle due to potential contamination of a cosmological sample of Type Ia supernovae.

1.2.3.1 Type Ia supernovae

Type Ia supernovae play a key role in cosmology as they are standardisable. Phillips (1993); Hamuy et al. (1996) demonstrate the positive correlation between

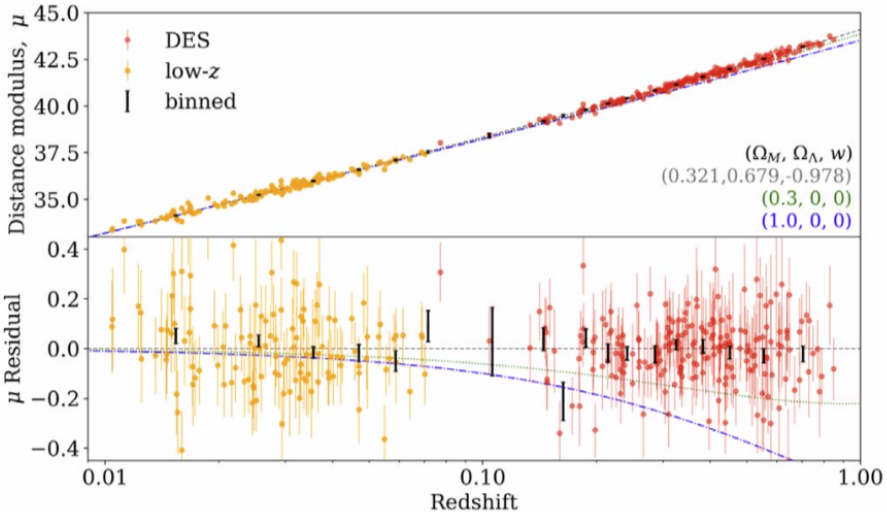


Figure 1.2: Top: the Hubble diagram constructed with the DES-SN3YR sample from Abbott et al. (2019). The grey line is the best fit model and the green are blue lines represent theoretical models considering no Λ component with $\Omega_m = 0.3$ and 1.0 respectively. Bottom: residuals to the best fit model.

their luminosity and light curve width. After applying corrections for light curve shape and colour (Tripp, 1998), and also host-galaxy properties (Murakami et al., 2021), they exhibit very similar peak luminosity. Combining this with redshift measurements, Type Ia supernovae therefore provide an excellent standardisable candle with which to measure the Universe's expansion. Their standardisable nature originates as stars exploding at the same mass. The Chandrasekhar mass ($1.44M_{\odot}$) was calculated as the theoretical limit at which a white dwarf can be supported by electron degeneracy pressure (Chandrasekhar, 1931). If a carbon-oxygen white dwarf accretes matter from a binary companion, as it approaches this mass limit it will reach the ignition temperature for carbon fusion and undergo a runaway thermonuclear explosion (Filippenko, 1997). Progenitors of such explosions are believed to be very similar, leading to supernovae of a characteristic brightness (Liddle, 2003).

Similar to Hubble's original galaxy recession diagram (Hubble, 1929), we can construct Hubble diagrams with any standard candle. Hubble diagrams with Type Ia supernovae are constructed as a function of redshift, where a distance modulus $\mu = 5\log_{10}(d_L/10 \text{ pc})$ is used as a proxy for distance, characterised by the

apparent magnitude m_B (at peak, in rest-frame B band) of Type Ia supernovae (Betoule et al., 2014):

$$\mu = m_B - (M_B - \alpha \times X_1 + \beta \times C) \tag{1.21}$$

where α, β and M_B[1] are nuisance parameters, and m_B, X_1 and C are light-curve parameters fit using a spectral sequencing model, typically SALT2 (Guy et al., 2007). Fig. 1.2 shows a Hubble diagram constructed with Type Ia supernovae from DES-SN3YR sample, including the binned and unbinned data.

At different redshifts we observe different parts of a source's spectrum. Applying a K-correction converts an observed magnitude into one that would be observed in the rest frame in another filter, in this case the B band (Hsiao et al., 2007; Oke & Sandage, 1968). For a source observed in filter R with apparent magnitude m_R, to transform into an absolute magnitude M_Q in another filter Q, the K-correction term K_{QR} is defined by

$$m_R = M_Q + \mu + K_{QR} \tag{1.22}$$

and is calculated by the following:

$$K_{QR} = -2.5\log_{10}\left[\frac{1}{1+z}\frac{\int \mathrm{d}\lambda_o \lambda_o L_\lambda \left(\frac{\lambda_o}{1+z}\right) R(\lambda_o) \int \mathrm{d}\lambda_e \lambda_e g_\lambda^Q(\lambda_e) Q(\lambda_e)}{\int \mathrm{d}\lambda_o \lambda_o g_\lambda^R(\lambda_o) R(\lambda_o) \int \mathrm{d}\lambda_e \lambda_e L_\lambda(\lambda_e) Q(\lambda_e)}\right], \tag{1.23}$$

where λ_o and λ_e are the observed and emitted wavelengths respectively, L_λ is the source luminosity, g_λ is the spectral flux density for the zero-magnitude source in the given filter, and $R(\lambda)$ and $Q(\lambda)$ are the mean contributions to the detector signal in their respective filters, i.e. the probability of a photon with wavelength λ being counted (Hogg et al., 2002).

From these diagrams we can parametrise the luminosity distance following a cosmological model, and constrain values in different cosmological models, e.g.,

[1] M_B is the *absolute magnitude*, i.e. the apparent magnitude at a distance of 10 pc, at peak.

Ω_m in a ΛCDM model, Ω_m and dark energy equation of state w in a flat, non-cosmological constant (wCDM) model, or Ω_m, w_0 and w_a in a flat, variable dark energy (w_0w_aCDM) model.

1.2.4 Best current measurements

Type Ia supernovae offer a parametrisation of the Universe's expansion up to redshifts of approximately $z \sim 1$ (Betoule et al., 2014). It becomes difficult to observe supernovae beyond this and there are few samples that reach further, although recent efforts have been able to observe Type Ia supernovae at redshifts as high as $z \sim 2.3$ using photometry from the Hubble Space Telescope (Hayden et al., 2021; Riess et al., 2018; Williams et al., 2020). For a more complete analysis, cosmological measurements are often combined from multiple independent methods.

1.2.4.1 The cosmic microwave background

The cosmic microwave background radiation is an almost perfect black body spectrum, with a precise estimation in temperature of $T_0 = 2.72548 \pm 0.00057$ K from the literature (Fixsen, 2009). It permeates the whole Universe and we can observe its *surface of last scattering* from when CMB photons decoupled from matter in the Early Universe[1], therefore providing a very high-redshift anchor ($z \sim 1100$) to cosmological constraints from supernovae. Jones & Wyse (1985) determined that, modelling a 'visibility function' that measures the probability of photon scattering within redshift dz, the last scattering surface is well-approximated by a Gaussian with mean $z = 1067$ and width $\Delta z \sim 80$.

Cosmology with the CMB comes from observation of anisotropies in its temperature, due to primordial density fluctuations. The origins of the temperature anisotropies found in the CMB are discussed in Hogan et al. (1982); White et al. (1994). To determine cosmological parameters from the CMB, the temperature fluctuations of each pair of points in the sky separated by angle θ are combined

[1]The existence of the CMB is further evidence for Big Bang cosmology as decoupling of light from matter occurred due to the Universe's cooling and decreasing density with expansion.

and described by the *angular power spectrum*, a spherical harmonic expansion of CMB temperatures. These fluctuations has previously been observed by the COBE (Boggess et al., 1992) and WMAP (Bennett et al., 2013) missions.

de Bernardis et al. (2000) provides an explanation of determining cosmological parameters from the angular power spectrum: The angular scale θ between fluctuations depends on multipoles l, where $\theta = \pi/l$. The position of peaks in the power spectrum therefore depends on multipole values. The Planck mission offers the latest and highest resolution measurements of CMB anisotropies. Constraints on cosmological parameters from seven acoustic peaks in the CMB are discussed in Planck Collaboration et al. (2020). The position of the first peak depends on the geometry of the Universe Ω_0, and observations suggest that the Universe is flat (where $l \sim 200$); combining with baryon acoustic oscillation (BAO; § 1.2.4.2) data, Planck Collaboration et al. (2020) constrains an energy density term for curvature to be $\Omega_K = 0.0007 \pm 0.0019$ ($\Omega_K = 0$ implies flat geometry). We can include energy density terms for both radiation $\Omega_r(1+z)^4$ and curvature $\Omega_K(1+z)^2$ into equation 1.20, although Planck data suggests both of these are negligible. Using the ΛCDM model of the Universe, results from Planck imply the density of matter to be $\Omega_m = 0.3158 \pm 0.0073$. Constraints are strengthened further to $\Omega_m = 0.3111 \pm 0.0056$ when combining with BAO measurements.

1.2.4.2 Baryon acoustic oscillations

Another independent method for constraining cosmological parameters comes from baryon acoustic oscillations, that can also be combined with both CMB and supernova data. The primordial density perturbations in the early Universe that are responsible for temperature fluctuations also give rise to BAO. Before the decoupling of light from matter, density perturbations in the dense plasma of electrons, baryons, dark matter and photons would produce sound waves, resulting in acoustic signatures that are seen much later in large-scale structure. Eisenstein et al. (2005) presents the first detections of the BAO signature. Using a spectroscopic sample of 46,748 galaxies between $z = 0.16$–0.47 from the Sloan Digital Sky Survey (SDSS), they find a characteristic peak in the redshift-space

correlation function ξ[1] at a comoving separation of approximately 150 Mpc (the BAO length scale, l_{BAO}) that matches predictions from theory.

The angles $\theta_{\rm BAO}$ between acoustic peaks provide a measure of the Universe's geometry and can be used to test cosmological models. The BAO angle relates the comoving length scale and the angular diameter distance d_A to the peak ($\theta_{BAO} = l_{BAO}/d_A$). Using spectroscopic redshift information, the given cosmological model can be numerically solved to determine parameters such as Ω_m and w. BAO therefore provides robust constraints on the accelerated expansion of the Universe as an independent probe to investigate the presence of dark energy (Blake & Glazebrook, 2003; Linder, 2003a).

Percival et al. (2010) provides a more up-to-date cosmological BAO study, comprising ∼900,000 SDSS galaxies, and Alam et al. (2017) provides a much more recent analysis, using a total sample of ∼1.2 million SDSS galaxies. Future BAO observations will come from the Euclid mission (Laureijs et al., 2011), which will map out large-scale structure covering the last 10 billion years.

1.2.4.3 Cosmology analyses with Type Ia supernovae

The most recent cosmological analyses using Type Ia supernovae are those from the Dark Energy Survey (3-year sample; DES-SN3YR, Abbott et al. 2019) and Pantheon (Scolnic et al., 2018b) samples.

Similar to the Hubble diagram (Fig. 1.2), cosmology parameters can be calculated assuming a model and solving numerically. Constraints on Ω_m and Ω_Λ, and Ω_m and w are shown in Figs. 1.3 and 1.4 respectively. Using a flat ΛCDM model, DES-SN3YR obtains $\Omega_m = 0.331 \pm 0.038$ (and therefore $\Omega_\Lambda = 0.669 \pm 0.038$). Combining with Planck CMB data, relaxing the cosmological constant assumption and assuming a flat geometry (the wCDM model), they find $\Omega_m = 0.321 \pm 0.018$ with $w = -0.978 \pm 0.059$. Finally, combining with CMB data and BAO data from three different studies, assuming a flat geometry with a variable dark energy equation of state ($w_0 w_a$CDM model) they find $\Omega_m = 0.316 \pm 0.011$, $w_0 = -0.885 \pm 0.114$ and $w_a = -0.387 \pm 0.430$.

[1] ξ is a measure of the probability that a galaxy will be found at a given comoving distance from another.

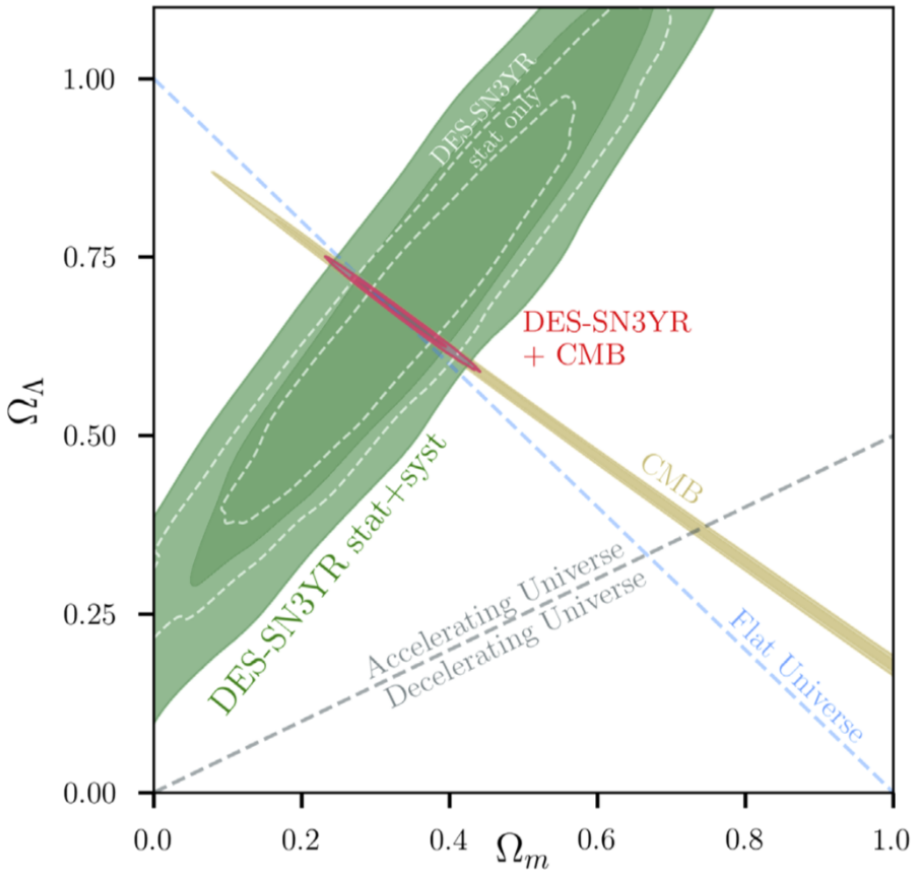


Figure 1.3: Inclusion of CMB data with the DES-SN3YR sample puts tight constraints on the densities of matter and Λ. This assumes a ΛCDM model and includes 68% and 95% confidence levels. Credit: Abbott et al. (2019).

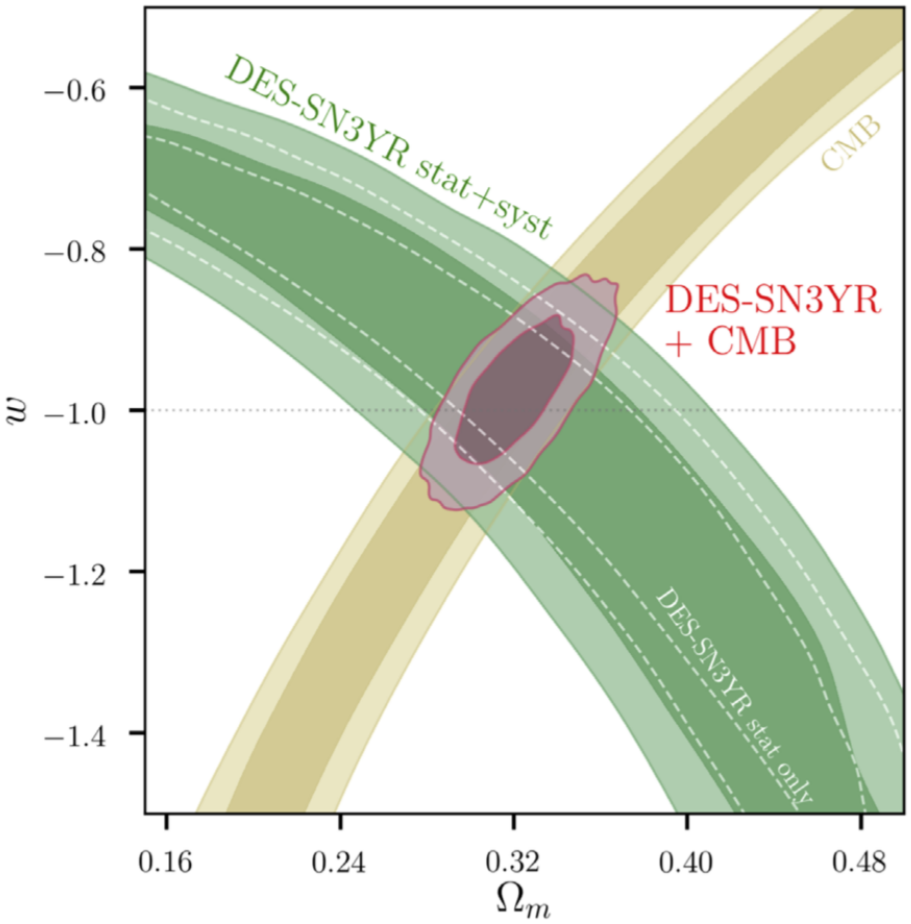


Figure 1.4: Constraints on the density of matter and dark energy equation of state using CMB data with the DES-SN3YR sample. This is relaxing the assumption of a cosmological constant (a wCDM model) and includes 68% and 95% confidence levels. Credit: Abbott et al. (2019).

Similarly, assuming a flat wCDM the Pantheon sample combined with Planck CMB data yields $\Omega_m = 0.307 \pm 0.012$ and $w = -1.026 \pm 0.041$. Combining the Pantheon sample and CMB measurements with BAO data, and assuming a flat $w_0 w_a$CDM model, they find $w_0 = -1.007 \pm 0.089$ and $w_a = -0.222 \pm 0.407$.

The consistency between different studies appears to indicate that, assuming a geometrically flat Universe, matter makes up approximately 30% of the total energy density, with dark energy making up the remaining 70%. Furthermore, the dark energy density is close to constant with expansion.

Chapter 2

Next Generation Cosmology

2.1 Upcoming ground-based astronomical facilities

In this chapter we discuss the innovative upcoming facilities that will be useful for Type Ia supernova cosmology and then we will introduce the surveys that are relevant to our research.

2.1.1 Vera C. Rubin Observatory

The Vera C. Rubin Observatory (Fig. 2.1) is currently being built at the summit of Cerro Pachón in the Northern Chilean Andes, one of the best sites in the world for astronomical observations. The local atmosphere is very dry, providing clear skies, and weather data from Cerro Tololo Inter-American Observatory (10 km from the Rubin site) suggests that observations can be conducted on >80% of nights, with a mean astronomical seeing of 0.67 arcsec in the g band (LSST Science Collaboration et al., 2009). It will be built among other facilities including Gemini-South and the Southern Astrophysical Research telescopes, where the exceptional quality of observational images has been proven. The Rubin Observatory's main function will be carrying out the Legacy Survey of Space and Time (LSST, §2.2.1), with operations currently expected to start around the end of 2023.

Figure 2.1: The Vera C. Rubin Observatory at the summit of Cerro Pachón, taken in September 2021. Credit: Bruno C. Quint, Rubin Obs/NSF/AURA lsst.org.

The Rubin Observatory will revolutionise astronomical sky surveys due to its large primary mirror (diameter of 8.4 m) and wide field of view (9.6 deg^2), but also in particular its immense data stream, gathering ∼20 TB of data per night and covering the visible night sky every 3–4 nights. With this however, comes the enormous task of finding the information that is useful for astronomers. Processing such a huge amount of data puts Rubin's output firmly in the scope of big data. Humans performing traditional data analysis methods will be insufficient in keeping up with the volume of data streaming from Rubin. Alert brokers are softwares being developed that receive information from the Rubin Observatory's difference imaging for identifying new astronomical sources, including transients. Once an alert of a new discovery is created, the broker distributes the data to the scientific community. During operations this will be the first step in the data pipeline to translate raw data into science[1].

The raw data will be images taken using the Rubin Observatory LSST Camera. This digital camera will be the largest ever constructed (1.65 m × 3 m), comprising 189 16-megapixel CCDs (charge-coupled devices – silicon-based detectors)[2]. Photometry will be determined by the camera's filters u, g, r, i, z and y, spanning the ultraviolet (330 nm) to near-infrared (1080 nm). Each filter has a wavelength range of approximately 100 nm. Multiple filters are used to make comparisons between flux measurements to help understand light curve characteristics. Exposure times for each image will be either 2 × 15 s or 1 × 30 s, which both have advantages and disadvantages for different science goals. Final choice will be decided in the commissioning phase of LSST (Lochner et al., 2021).

The Rubin Observatory will be a cornerstone of next-generation astronomy and will be such an advancement that data-mining techniques and artificial intelligence will be instrumental in taking our understanding of astrophysics and cosmology to an unprecedented level.

Figure 2.2: The hexagonal array of 4MOST's 2436 fibres. Credit: Joe Liske, 4most.eu.

2.1.2 4MOST

Whilst Rubin will provide the photometry for astronomical observations in the coming years, a deeper understanding can be established with spectroscopy. The 4-metre Multi-object Spectroscopic Telescope[1] (4MOST) is an instrument under construction by the European Southern Observatory[2] (ESO). It will be installed on the Visible and Infrared Survey Telescope for Astronomy[3] (VISTA) in Chile, at a similar latitude to the Rubin Observatory and is expected to have its first light at a similar time.

For at least 5 years, VISTA will be dedicated purely to 4MOST, in which observing time will be split between consortium (70%) and community (30%) surveys. There are 10 consortium surveys, including 5 galactic and 5 extragalactic surveys. Survey 10, the Time-Domain Extragalactic Survey (TiDES §2.2.2), will be conducting a follow-up campaign for transient events discovered by the Rubin Observatory's LSST survey.

The 4MOST facility will be able to operate multiple astronomical surveys simultaneously thanks to its unique multi-plex fibre system. The 4MOST instrument comprises 2436 individual fibres, that split into two low-resolution spectrographs and one high-resolution spectrograph. This multitude of fibres means that in a given 4MOST pointing (a hexagonal 4.2 deg^2 field of view), each fibre can be directed towards their own individual targets. Fibre positioning is controlled by the AESOP fibre positioner (Brzeski et al. 2018, Fig. 2.2) and takes less than 2 minutes. The accuracy of fibre positioning is expected to be better than 0.2 arcseconds (de Jong et al., 2019). 4MOST's wavelength coverage is 370–950 nm and it has a spectral resolving power of 4,000–21,000.

Spectroscopy exposures typically take much longer than those for photometry as the incoming light is dispersed; it takes time for a spectrum to reach a high-enough signal-to-noise ratio (SNR) to be considered useful, depending on the scientific objective. Current 4MOST simulations combine observing fields of the

[1] https://www.lsst.org/scientists/alert-brokers
[2] https://www.lsst.org/about/camera
[1] https://www.4most.eu/cms/
[2] https://www.eso.org/public/
[3] https://www.eso.org/sci/facilities/paranal/telescopes/vista.html

same sky coordinates and instrument position angle into observing blocks (OBs, Tempel et al. 2020a) and the duration of the OBs are limited to a total exposure time of 1 h to follow ESO's scheduling constraints. If objects are too faint, then the 1 h exposure will not be long enough to reach a desired SNR. The 4MOST exposure time limit therefore creates a magnitude limit for observations, depending on a SNR criterion.

2.1.3 Extremely Large Telescope

ESO has begun construction of the Extremely Large Telescope (ELT) which will be the world's largest optical and near infrared telescope, with a 39 m diameter primary mirror[1]. To complement its immense size, it will operate with an adaptive optics system in order to ameliorate the effects of astronomical seeing on observations and produce sharp images with very high resolution over a 1 arcminute^2 field of view (Ciliegi et al., 2021). The ELT will join its many other ESO cohorts in Chile. It is being built on Cerro Armazones, not far from Cerro Paranal, the site of the Very Large Telescope[2] (VLT).

The ELT will be used to implement a wide range of scientific endeavours, using both imaging and spectroscopy. ELT science will extend from local solar system observations, to exoplanets and stellar and galactic physics within the Milky Way, and beyond to extragalactic sources and into the early Universe. Observations of the most distant sources will include Type Ia supernovae for probing cosmology and some of our most fundamental understanding of physics. The ELT will provide extension of the Type Ia supernova Hubble diagram to currently unexplored redshifts, beyond $z = 2$ and up to $z \approx 4$ (Hook, 2013).

2.2 Upcoming surveys

With these new facilities and instruments, the organisation of putting the science objectives into practice requires formulation of telescope surveys. For dark energy

[1] https://elt.eso.org/telescope/

[2] https://www.eso.org/public/unitedkingdom/teles-instr/paranal-observatory/vlt/

science, we discuss some of the frontrunners in next generation surveys.

2.2.1 LSST

The Rubin Observatory's main mission during its first 10 years of operations will be the Legacy Survey of Space and Time (LSST[1]). As mentioned in §2.1.1, this survey will be revolutionary for astronomy due to its immense intake of data. To accommodate the large scope of science in LSST, survey strategy and cadence optimisation is a highly non-trivial task and details are still being finalised (LSST Science Collaboration et al., 2017). LSST's primary focus will be its main, wide-fast-deep (WFD) survey, which will cover ∼18,000 deg^2 every 3–4 nights. In the *ugrizy* bands, the 5σ single-visit depths are 23.9, 25.0, 24.7, 24.0, 23.3, 22.1 (AB magnitudes) respectively (Ivezić et al., 2019). For supernova discovery, the WFD survey will reach up to $z \sim 0.8$. LSST's deep drilling fields (DDFs) are at least four fields covering tens of square degrees that will be visited with a much higher cadence to reach deeper coadded magnitudes (∼10× fainter, LSST Science Collaboration et al. 2009). These will yield well-sampled supernovae peaking around $z \sim 0.7$ and reaching beyond $z = 1$ (Ivezić et al., 2019).

The Dark Energy Science Collaboration[2] (DESC) is the division of LSST that is focused on exploring the nature of dark energy. DESC itself is also split into its own working groups: Weak Lensing; Large Scale Structure; Time Domain; Clusters; Modelling and Combined Probes; Photometric Redshifts; Observing Strategy; External Synergies; Dark Matter. Our work on photometric classification of supernovae is conducted as part of the Time Domain working group.

2.2.2 TiDES

The Time-Domain Extragalactic Survey (TiDES) is one of ten 4MOST consortium surveys (de Jong et al., 2019; Swann et al., 2019). It will be dedicated to spectroscopic follow-up of transients for cosmology. TiDES is split into three sub-surveys: (i) TiDES-SN, which will focus on spectroscopic classification of live transients;

[1]The acronym LSST was formerly the Large Synoptic Survey Telescope, encompassing both the observatory and survey.

[2]https://lsstdesc.org/

(ii) TiDES-Hosts, which will provide spectroscopic redshifts of supernova host galaxies; (iii) TiDES-RM, which will produce reverberation mapping of active galactic nuclei (AGN) from DDF observations.

Over the 5 year survey, TiDES will accumulate >30,000 live transient spectra[1]. These will reach magnitudes as faint as $r_{\rm AB} \approx 22.5$ mag, given by TiDES' SNR criterion, discussed in detail in § 4.4. Reverberation mapping of at least 700–1,000 AGN from repeat observations in deep drilling fields will provide a complementary cosmological analysis on the Hubble diagram, extending to redshifts as high as $z \approx 2.5$. Host-galaxy spectroscopic redshifts are used to plot Type Ia supernovae on the Hubble diagram. The final cosmological sample will not only be Type Ia supernovae from TiDES-SN, but will also include those identified from photometric classification by using the TiDES-SN sample as a spectroscopic training sample. This means that even those LSST transients without live spectra, but with host-galaxy redshifts, can be included in a cosmological analysis after being identified as Type Ia by photometric classification. The study Mitra & Linder (2021) finds that robust supernova cosmology requires spectroscopic redshifts, particularly at redshifts $z \lesssim 0.2$–0.3 where photometric redshifts are not reliable. Graham et al. (2020) also finds that 10% of galaxy photometric redshifts from LSST photometry will be outliers at $z = 0.5$, reaching even higher percentages at lower redshifts (where outliers are those with redshift error greater than 3 times the robust standard deviation, or 0.06). This is true of host galaxies of any type of object, both Ia and non-Ia (including potential contaminants) and will be relevant later in § 3.5 where we discuss results of photometric classification using different redshift estimates. TiDES will ensure a large sample (70,000) of reliable spectroscopic host-galaxy redshifts, enabling the largest cosmological sample of Type Ia supernovae to date, by at least an order of magnitude.

To achieve its scientific goals, TiDES will be using its allocated 250,000 fibre-hours on 4MOST. TiDES is different from other 4MOST consortium surveys as it will not be driving the pointings of 4MOST, but will instead be exploiting the fact that wherever 4MOST points in the sky there will be recently discovered LSST transients to follow-up. TiDES will be 'piggy-backing' on the other consortium

[1]The sample numbers have been provided through private communication in TiDES.

surveys as the target density of transients is not high enough for efficient observations on its own; TiDES utilises approximately 2% of 4MOST fibres (30–35 low-resolution spectrograph fibres), so it would not be efficient to use 4MOST exclusively for LSST transients. There will be a rapid turnaround time of 3–4 days in which to target the allocated fibres onto objects identified from LSST transient alert brokers.

Chapter 3

Photometric Classification of Supernovae

With the many Type Ia supernova discoveries from LSST, we will be able to test cosmological models and constrain parameters such as the dark energy equation of state to a much higher degree of precision than from any previous dataset. To use supernovae as cosmological probes, we first need to be sure that they are in fact Type Ia. Supernova type is traditionally determined by the chemical signatures that appear in their spectra, for example the presence of silicon in Type Ia supernovae (Filippenko, 1997; Walker et al., 2010), however, getting spectra of all LSST transients is not realistic due to expensive spectroscopic resources. Not wanting to waste the potential supernova science of all these objects, we therefore need to consider other methods of classification for the transient events that are not spectroscopically followed-up. Hence, photometric classification of supernovae using machine learning provides a solution.

Photometric classification with machine learning is a process that takes supernova light curve observations, generally with multiple filters, and determines their types based on information learnt from a given training sample of supernova light curves with confirmed type. In preparation for LSST and other future surveys, there has recently been a great focus into what makes a good training sample for photometric classification of supernovae. As with many typical machine learning

problems, a training sample that is representative of the whole dataset that is to be classified – the 'test set' or 'target sample' – seems a necessity (Lochner et al. 2016; Charnock & Moss 2017; Ishida et al. 2019; Muthukrishna et al. 2019a; Möller & de Boissière 2020). A representative training sample is one whose feature-space distributions are similar to those of the test set. Machine-learning models trained on samples that are representative of the target distribution are expected to perform well in classification tasks, so long as they have sufficient coverage of the test data (Beck et al., 2017). There are broad variations in light curves and supernovae have a wide range of magnitudes and redshifts. A representative training sample should include the features associated with these variations.

None the less, works into data augmentation methods show that focusing on accumulating a spectroscopic sample of supernovae that is fully representative may not be necessary. As long as one starts with a sample that has reasonable coverage of the full test set, augmentation can fill the gaps to create a much more representative training sample. Using Gaussian processes to model supernova light curves, it is possible to create new simulated light curves that cover more of the test set feature space and add them into the training sample, making it artificially more representative. This approach is used in the works by Revsbech et al. (2018) and Boone (2019), yielding very promising classification results. The latter of these was the winning solution to the Photometric Light Curve Astronomical Time-Series Classification Challenge[1] (PLAsTiCC; results of the challenge are discussed in Hložek et al. 2020), which required classifying simulated LSST data using a provided non-representative training sample. The training sample mimicked a real set of light curves (of many types of object, not just supernovae) with spectroscopically-confirmed type and a preference to brighter, low-redshift objects. With augmentation to create artificial light curves and help cover the whole feature space, less time is required from spectroscopic resources to build a faint training sample.

We also consider the role of redshift in the photometric classification of supernovae. For Type Ia cosmology we require spectroscopic redshifts of supernovae, as cosmology with photometric redshifts will be skewed and is prone to contamination (Linder & Mitra, 2019; Mitra & Linder, 2021). At the end of the TiDES survey, we

[1] https://www.kaggle.com/c/PLAsTiCC-2018

will have a spectroscopically confirmed sample of supernovae that will be used as the basis of our training sample. We will also have spectroscopic redshifts for many host galaxies of LSST supernovae for which we do not have a classification. These are the supernovae that we will want to photometrically classify for cosmology. Spectroscopic redshifts are necessary for cosmology, but can also be used as an additional feature in our classifiers. Lochner et al. (2016) concluded that including photometric redshifts of supernova host galaxies does not have a significant impact on classification when using representative training samples, although the level of accuracy is model- and algorithm-dependent. We investigate the three cases of using spectroscopic, photometric and no redshift in classification.

In this research, we investigate the potential success of photometric classification using machine learning by simulating a realistic training sample. The training sample will be the TiDES spectroscopic supernova sample, which is discussed in Chapter 4.

3.1 Choice of software

Before we could test photometric classification of transients, we had to choose an appropriate software to use. At the start of the project we decided to use SNMACHINE (Lochner et al., 2016), a classification pipeline using machine-learning algorithms that is available through the Rubin Observatory LSST Dark Energy Science Collaboration[1] (DESC). This was chosen because SNMACHINE matches our goals as it is designed for photometric classification of supernovae in the Dark Energy Survey (DES) and LSST, and was readily available at the time.

In recent years there have been a myriad of photometric classifiers created. SUPERNNOVA (Möller & de Boissière, 2020) includes a deep recurrent neural network architecture for transient classification. Following our main investigation, we applied this software to a simulated LSST dataset to investigate contamination in the spectroscopic sample, discussed in Chapter 6.

[1] https://lsstdesc.org/

Another notable software example is ActSNClass (Ishida et al., 2019) a constituent of the COINtoolbox[1]. This framework applies active learning (Cohn et al., 1996), a method to statistically select the most appropriate objects for optimal training. While blanket targeting of all possible transients for spectroscopic follow-up will produce an unbiased sample, active learning offers a way to plan follow-up strategy to create an optimal training sample using limited spectroscopic resources.

Other examples include PELICAN (Pasquet et al., 2019), SCONE (Qu et al., 2021) and RAPID (Muthukrishna et al., 2019a).

3.1.1 Dataset

For our simulations we use the Supernova Photometric Classification Challenge (SPCC) dataset (we use the simulations that were updated following the original challenge, Kessler et al. 2010a,b). The data are simulated light curves of 21,319 supernovae of different types: Ia, Ib, Ic, Ibc, II, IIP, IIL and IIn. Later in § 5.4 we discuss the balance of classes and how they are affected by different types of training sample, with respect to the test set.

The SPCC light curves have been simulated to mimic DES observations, using the filters g, r, i and z. LSST has additional filters u and y, which may improve classification, although is very close to the SPCC as LSST's supernova cosmology focus will be on the g, r, i, z bands (The LSST Dark Energy Science Collaboration et al. 2018 finds that filters u and y provide negligible cosmological information), with a similar cadence of observations every few days in each filter. The light curves consist of flux measurements and associated uncertainties in the four bands at times specified by the Modified Julian Date. In snmachine, the light curves are aligned such that they all start at time $t = 0$.

In this work we primarily consider the binary classification of Type Ia vs. non-Ia (which we define as the positive class vs. negative class), due to our focus on applications to Type Ia cosmology. However, we also run a few tests in which snmachine returns a classification probability for each supernova being either a

[1] https://github.com/COINtoolbox; the COIN (COsmostatistics INitiative) collaboration is an international community focused on data-driven astronomy research

Type Ia, Ibc (which includes types Ib, Ic, Ibc) or II (which includes types II, IIP, IIL, IIn), as in, e.g. Möller & de Boissière (2020) and also the many solutions to the SPCC and PLAsTiCC challenges. In this case, we still apply a binary Ia vs. non-Ia classification, but with the aim of investigating whether considering Type Ibc and Type II light curves separately in the training would reduce the number of false positives (non-Ia light curves classified as Ia).

3.1.2 Feature extraction

The process for classification starts with extracting features from all the supernova light curves in the dataset. We use the wavelet decomposition method implemented in SNMACHINE that extracts the wavelet coefficients that parametrise the light curves of each supernova using a Gaussian process regression. An advantage of this feature extraction method is that it requires no prior information about supernova light curves. It is therefore independent of any physical assumptions and is a purely mathematical framework.

For extraction of wavelet features, each supernova light curve is first interpolated onto a uniform grid of 100 points and each point in time is associated with a Gaussian distribution. The Gaussian process generalises across all distributions, encompassing the mean and covariance functions between points (MacKay, 2003). The mean function and covariance matrix hyperparameters are determined using `GaPP` (the Gaussian process regression software used in Seikel et al. 2012). Next, a two-level wavelet transform decomposes the light curve into a set of wavelets. This results in a highly redundant 1,600 (400 per filter) wavelet coefficients per supernova. The wavelet coefficients define the position of a wavelet basis (shape of a specific type of wavelet) on the light curve.

To reduce the dimensionality whilst preserving the useful information, a principal component analysis (PCA, Hotelling 1933; Pearson 1901) is applied to the wavelet coefficients. PCA derives the eigenvalues of the correlation matrix that represent the variability of each feature. Large eigenvalues imply high levels of variability. Therefore, features with negligible eigenvalues are discarded. After PCA, there are 20 features per light curve. This was chosen as Lochner et al. (2016) finds that reducing the number of features from 1,600 to 20 using PCA

retains 98% of the dataset's information (by weighting the significance of features based on their eigenvalues). If redshift is included as an additional feature, we add this to the feature set for each supernova, making a total of 21 features.

Feature extraction is the main computational bottleneck in the SNMACHINE pipeline. However, after extracting light curve features from a dataset, e.g. SPCC, they can be saved and used again to avoid extracting features every run. When introducing new dataset objects, e.g. augmented supernovae, features must be extracted with every new training sample. This means that the pipeline is much more efficient using training samples that are constructed from only the original dataset.

Assuming that photometric or spectroscopic redshift information is available, this can be included in the table of light curve features independently following the wavelet feature extraction. Hence, the same saved set of wavelet features can be used in all tests using the same dataset (e.g. original, or original plus augmented objects), and if necessary, can be merged with redshift information prior to the machine learning.

The whole classification pipeline was run on the Lancaster University High End Computing (HEC) cluster. Feature extraction was run in parallel across many CPUs (Central Processing Units; 16 CPUs on one core) to speed up computation. To allocate optimal computing resources, the training and classification stages were run separately on a single CPU.

3.1.3 Machine-learning algorithms

SNMACHINE's machine-learning classification algorithms are trained to associate feature values with the chosen classes (e.g. Ia vs. non-Ia) from supernovae in the training sample. When presented with the test set light curves, SNMACHINE returns a probability of each supernova being either Type Ia or non-Ia. The classification algorithms are k-nearest neighbours (KNN), support vector machines (SVM), artificial neural networks (ANN) and boosted decision trees (BDT), and they are constructed from the SKLEARN implementations (Pedregosa et al., 2011)[1]. For the case of wavelet decomposition feature extraction, SNMACHINE's naive

[1] https://scikit-learn.org/stable/

Bayes algorithm performs barely better than randomly in classification (even in the case of a representative training sample) and is therefore disregarded.

3.1.3.1 Hyperparameters

Rather than choosing an arbitrary set of hyperparameters for each algorithm or running our own validation tests, we run the SNMACHINE pipeline with its built-in classifier optimisation method. The training sample is split into five 'folds', comprising four separate training subsamples and a validation sample. This follows k-fold cross-validation (note that this is not the same 'k' as in KNN) from Kohavi (1995). From a given range, the final hyperparameters that are used in classification of the test set are chosen in a grid search algorithm as the ones that maximise the validation sample's AUC scores. We will next spend some time describing each algorithm and the hyperparameters considered.

3.1.3.2 k-nearest neighbours

KNN is a clustering algorithm that assigns classifications based on its k-nearest neighbours in feature space (Altman, 1992). Using `sklearn.neighbors.KNeighborsClassifer`, the only hyperparameter allowed to vary was the number of neighbours (`n_neighbors`) from 1 to 176 in steps of 5. The `weights` parameter was chosen as 'distance', in which points are weighted by the inverse of their Euclidean distance. Other parameters were taken as their default values. The probability of an object belonging to a given class is calculated as the sum of neighbours' weights of that class divided by the sum all neighbours' weights. Classification of an object is then the class with the highest probability. For this algorithm, it is obvious why a representative training sample is necessary; if there are no close training neighbours to test set objects in feature space, it is unlikely that the algorithm can easily determine accurate classification.

3.1.3.3 Support vector machine

SVM is an algorithm that attempts to find the hyperplane that best separates classes in feature space (Cortes & Vapnik, 1995). The 'support vectors' are the

object vectors that define the margin that creates the largest separation between classes, and therefore generalises classification best. Using `sklearn.svm.SVC`, the `C` regularisation parameter is varied between 10^{-2} to 10^{5} including 5 values equally separated in log space. By the use of a kernel function (Aizerman et al., 1964), complex, high-dimensional datasets can be classified with SVMs by transforming the feature space into one in which linear separation is possible (Géron, 2017). This is known as the 'kernel trick'. `gamma` is a fine-tuning shape parameter of the kernel, in this case a Gaussian radial basis function (RBF), and is also varied on a log scale, including 5 values between 10^{-8} to 10^{3}. Other parameters used are the function's default values. After training the SVM model to separate classes, class probabilities of test objects are obtained by Platt scaling (Platt, 1999), a logistic regression of the SVM's scores that are fit by additional cross-validation on the training sample data.

3.1.3.4 Artificial neural network

ANNs are algorithms inspired by biological neural networks. Their versatility in how they are constructed (e.g. number of layers and neurons[1], convolutional and fully connected layers, etc.) makes them highly prevalent in both machine learning and deep learning tasks. In our case, the ANN takes the input features and maps them to output classes. The ANN is trained by back propagation, in which each neuron's weights are updated in order to minimise the loss function, where the loss function represents how erroneous the output is. The architecture used in this case is very simple, including only one fully-connected hidden layer, and created with `sklearn.neural_network.MLPClassifier`. The size of this layer (`hidden_layer_sizes`) is the only hyperparameter allowed to vary to optimise this algorithm, with its single layer going from 80 to 115 neurons in steps of 5. MLP stands for Multi-Layer Perceptron, used to describe this very simple form of input-to-output network and based on *threshold logic unit* neurons (Rosenblatt, 1960), where each input connection has an associated weight (Géron, 2017). The neuron's output is the weighted sum of its inputs (in our case the light curve features)

[1]Neurons are the connecting nodes in a network. They take input values, multiplying them by weights, summing them up and finally applying an activation function. This results in a single output that then connects to another layer's nodes or output classes.

applied to a hyperbolic tan activation function (the `activation` parameter). By normalising the activation function output values so that they sum to one, they effectively represent probabilities of an object belonging to the given classes.

3.1.3.5 Boosted decision tree

A BDT is an ensemble algorithm comprising many decision trees (Friedman, 2002). A decision tree algorithm maps input features to output classes through a series of decisions, i.e. testing whether feature values fall within particular ranges. These algorithms are simple and their decisions are easy to interpret, often being described as *white box* models, as opposed to *black box* models that are commonly associated with neural networks (Géron, 2017). On their own, decision trees can develop robust models on training data, although do not generalise well to other datasets, causing overfitting. Boosting is an ensemble method that is used to improve decision tree classifiers in which multiple models are combined into a stronger one. Boosting trains the same model on the same data multiple times, trying to improve at each iteration by focusing on incorrectly classified objects. The `sklearn.ensemble.AdaBoostClassifier` (Freund & Schapire, 1997) is used with a `base_estimator` defined by a `DecisionTreeClassifier` object with `entropy` (the `criterion` parameter measuring information gain) and `min_samples_leaf` between 5 and 45 in steps of 10. This final parameter gives the minimum number of samples required to be at a leaf node, where leaf nodes represent class labels, i.e. it is a limit to how far the samples can be split in a decision tree. In a decision tree, the probability of an object belonging to a given class is proportional to the fraction of trained objects of that class on the corresponding leaf node. Applying the ensemble method, the final probability is an average of all the probabilities of each decision tree. The other hyperparameter allowed to vary is `n_estimators` which gives the maximum number of estimators at which boosting is terminated, with values between 5 and 75 in steps of 10. An alternative ensemble method is a random forest, in which the same training algorithm is used on random subsets of the training data. In *bagging* (bootstrap aggregating), subsets are sampled with replacement (objects may appear more than once). The alternative to this is *pasting* where there is no replacement.

3.2 Classification performance and metrics

To assess the levels of success in photometric classification, we need to choose specific metrics to optimise. Our choice of metrics is determined by the problems we are considering. When looking at the overall performance of classification algorithms, we refer to their Receiver Operator Characteristic (ROC) area-under-the-curve (AUC) parameter, a common tool used with binary classification. ROC curves compare the True Positive Rate (TPR, a.k.a. completeness, efficiency) against False Positive Rate (FPR, a.k.a. contamination) for a range of probability thresholds, i.e. the threshold that the algorithm requires to apply the Type Ia classification to a supernova (the 'positive' class). TPR and FPR are defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{3.1}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \tag{3.2}$$

where TP is the number of true positives (Ia classified as Ia), FP is the number of false positives (non-Ia classified as Ia), TN is the number of true negatives (non-Ia classified as non-Ia) and FN is the number of false negatives (Ia classified as non-Ia).

For each run with SNMACHINE we produced ROC curves for each machine-learning algorithm. A ROC curve's AUC value equals 1 for a perfect classifier (TPR = 1 and FPR = 0) and 0.5 for a completely random classifier. However, high AUC does not reveal the full story and is not necessarily indicative of 'good' classification.

Given the large scope of objects to be observed by LSST, there may be the caveat of a small, 'strange' population of objects, e.g. superluminous supernovae, which could be completely misclassified. In the context of using Type Ia supernovae for cosmology, it is crucial that our classified sample has very low contamination and so we also consider purity as an essential metric. Purity is defined as

$$\mathrm{Purity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{3.3}$$

In general, there is a trade-off between TPR and purity.

For any classification problem, the measure of success depends on the choice of metric. For increasingly large datasets, e.g. from LSST, there will come a point at which systematic error dominates over statistical error[1]. Therefore, we assume that our classified sample is above the completeness level at which contamination from systematic effects dominates statistical error and we set a high target purity value of 95%. An in-depth look into when exactly this occurs for LSST requires further studies.

3.3 Visualising feature space – t-SNE

After extracting features from a dataset, it is useful to see how these features are related, especially with respect to the different classes and across training and test samples. In SNMACHINE, 21 features are extracted per supernova light curve (20 wavelet features plus redshift). An inspection between each supernova's set of features is therefore highly non-trivial. Visualisation of high-dimensional data is possible, however, if each datapoint is reduced to a 2D or 3D representation of feature space. This is achieved by adopting t-distributed stochastic neighbour embedding (t-SNE, Van der Maaten & Hinton, 2008), a method which clusters similar high-dimensional objects together. In this work we look at 2D representations of the 21-dimension feature space describing each supernova light curve. The t-SNE method shares a lot of similarities with both the k-nearest neighbours algorithm and PCA, namely a similarity calculation based on distances between points in high-dimensional feature space, and reducing dimensionality to a comprehendible level.

t-SNE calculates the pairwise Euclidean distance, giving the probability[2] of similarity between each pair of objects. To create a low-dimensional representation of the feature space, the 2D values that preserve these probabilities are determined[3].

[1] Statistical error increases by $\sqrt{N}$, where N is the number of objects in the dataset, whereas contamination rate caused by systematic error is proportional to N.

[2] This is the conditional probability that one object would pick another as its neighbour, assuming a Gaussian centered on the object.

[3] The units of these values, referred to as t-SNE X and Y, are arbitrary.

This is done by minimising the sum of Kullback-Leibler divergences (Kullback & Leibler, 1951) between the joint probabilities in the high- and low-dimensional spaces over all data points and using a gradient descent method. Generally, the closer together the objects are in a plot of the 2D values, the more similar they are in nature. Hence, clear separation between classes is indicative of intrinsic differences in their respective features and suggests that accurate classification is possible. It should be highlighted that these plots are for visual purposes only, and do not have any influence on actual classification.

In our t-SNE plots, we use `sklearn.manifold.TSNE` with the following parameter values: `perplexity` = 200 (related to the number of nearest neighbours); `n_iter` = 1000 (maximum number of iterations); `learning_rate` = 200 (for tuning the rate of gradient descent). Other parameters are their default values. This selection of (seemingly arbitrary) values was chosen by inspecting t-SNE plots of many different values and finding that this combination produces clear results.

3.4 Representative training sample

Before we explore photometric classification with our 4MOST training sample, we first follow the procedure from Lochner et al. (2016) to demonstrate what is possible when using representative training. First, we discuss what we mean by 'representative'.

In a given dataset with well-defined classes, a randomly drawn training sample of sufficient size has proportions of different supernova types equal to those in the test set (we discuss class balance in § 5.4). It is blind to supernova light curve parameters and has similar distributions in magnitude and redshift, shown in Figs. 3.1 and 3.2. Consequently, a randomly drawn training set samples the full range of feature values existing in the test set. To illustrate this, we show a two-dimensional representation of the 21 wavelet features (after PCA and including spectroscopic redshift), separated into training and test sets, and also by type (Ia vs. non-Ia). Shown in Fig. 3.3 for a randomly drawn training sample, training and test Type Ia supernovae occupy the same feature space, and similarly for training and test non-Ia supernovae. Hence, given sufficient size, a randomly

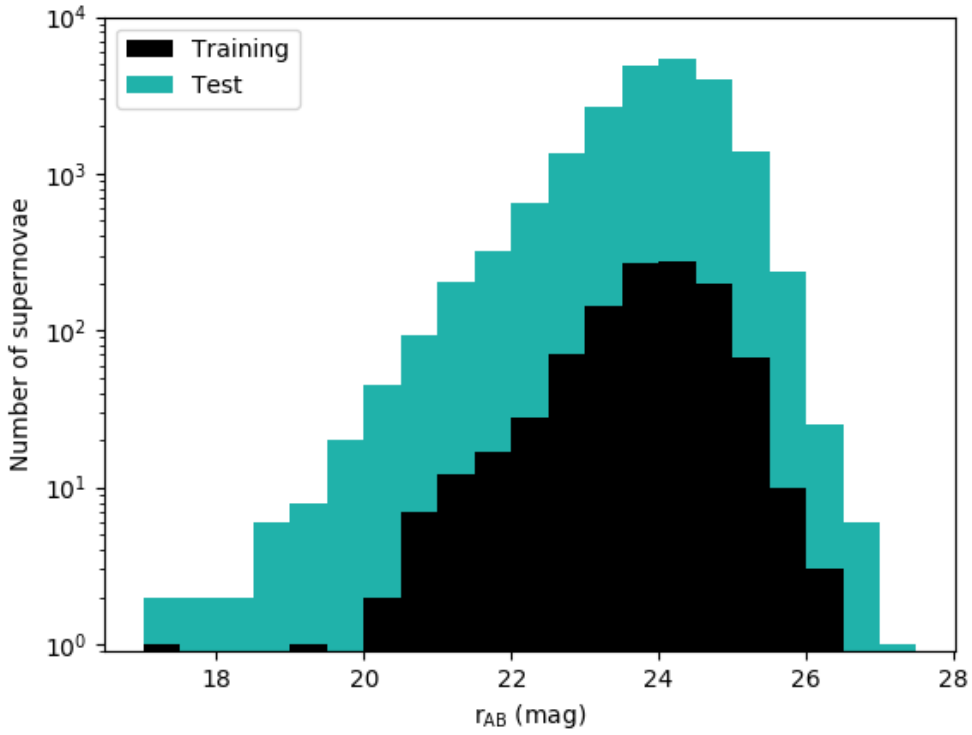


Figure 3.1: Stacked magnitude histogram of a random sample of 1,103 training supernovae and the corresponding test set.

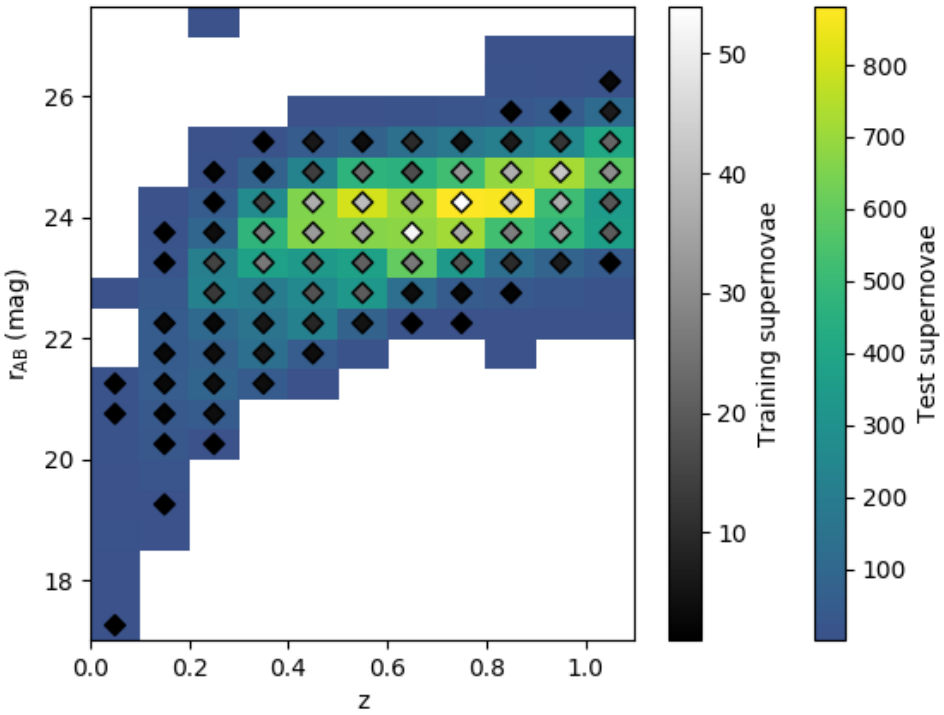


Figure 3.2: 2D histogram of the relative distributions of redshift and magnitude in a random sample of 1,103 training supernovae and the corresponding test set. Note that the bin containing a single faint supernova ($r_{\mathrm{AB}} > 27.5$), which appears anomalous to the rest of the dataset, is a result of this particular simulated light curve only having two very faint observations in the r-band.

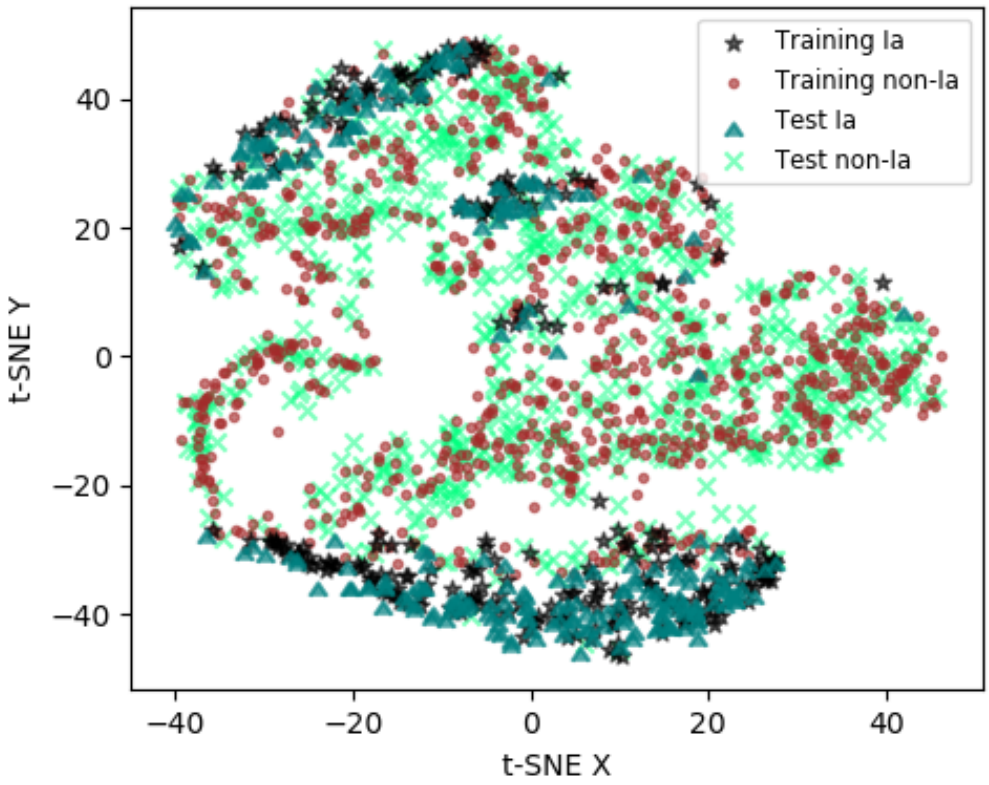


Figure 3.3: A t-SNE 2D representation of the 21-dimensional feature space after PCA and including spectroscopic redshift. Type Ia and non-Ia supernovae are found in their own respective clusters and regions of the plot. A randomly drawn training sample has supernovae of the same types occupying the same feature space as those in the corresponding test set. This plot only includes one tenth of the test set for clarity.

drawn training sample can be considered to be representative of the corresponding test, and, for the rest of this work, we therefore refer to a randomly selected training sample as being representative, as in Lochner et al. (2016).

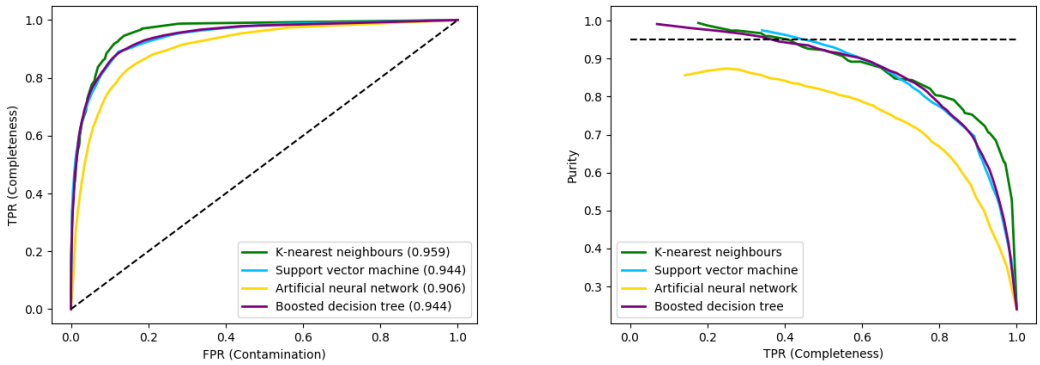
3.5 Results using representative training

In our tests, representative training samples are created by taking a random selection of 1,103 objects from the SPCC. This is the same size as the original sample in the classification challenge. We compare using the same training sample in an individual run, but with either the 'true' redshift[1], a photometric redshift or no redshift information used in both training and test samples to investigate which case is most successful for classification. The 'true' redshift is used to mimic a spectroscopic redshift and is defined as such from this point onwards. We do this for 20 runs and present the ROC curves and TPR-purity relationship for a typical example in Fig. 3.4.

To assess classification performance, we refer to ROC and purity curves throughout this work. The ROC curves' AUC scores are shown in brackets for each algorithm. The diagonal dashed lines in the ROC curve plots represent the case for a completely random classifier. The ANN algorithm is outperformed by the other algorithms, which all have comparably higher AUC scores and manage to reach our target purity of 95% (shown by the horizontal dashed line) with TPR $\approx$ 0.4–0.6 in all three redshift scenarios. Our ROC curves pass through the two theoretical classification extremes: $(\mathrm{TPR}, \mathrm{FPR}) = (0, 0)$, in which everything is classified as non-Ia, and $(\mathrm{TPR}, \mathrm{FPR}) = (1, 1)$, in which everything is classified as Ia. It should be noted that our AUC scores are calculated using only TPR and FPR values from classification based on the used range of probability thresholds. If TP = FP = 0, then the purity is undefined. In these cases, the purity curve may not start at TPR = 0. This also occurs if the minimum TPR value from our range of probability thresholds is non-zero, as purity is undefined below this TPR.

The resulting AUC scores for all runs with representative training are shown as boxplots in Fig. 3.5 and summarised in Table 3.1. For all three redshift scenarios

[1]This is the `SIM_REDSHIFT` parameter in the header of each supernova file.



No-z

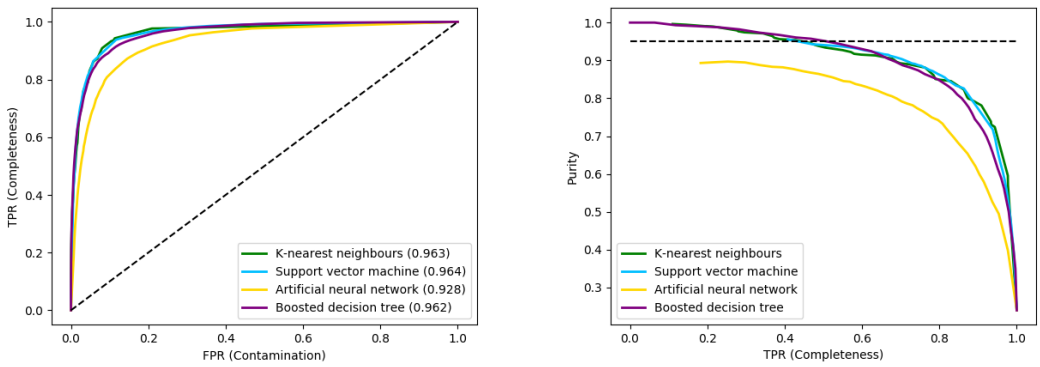
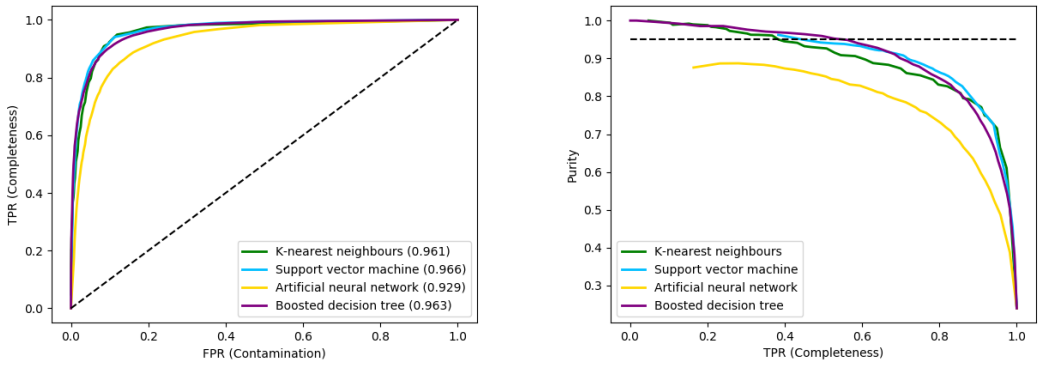


Photo-z



Spec-z

Figure 3.4: Results for the use of redshift for the same representative training sample, a typical example out of 20 runs. In the left column are the ROC curves, comparing TPRs and FPRs. In the right column are plots comparing purity with TPR.

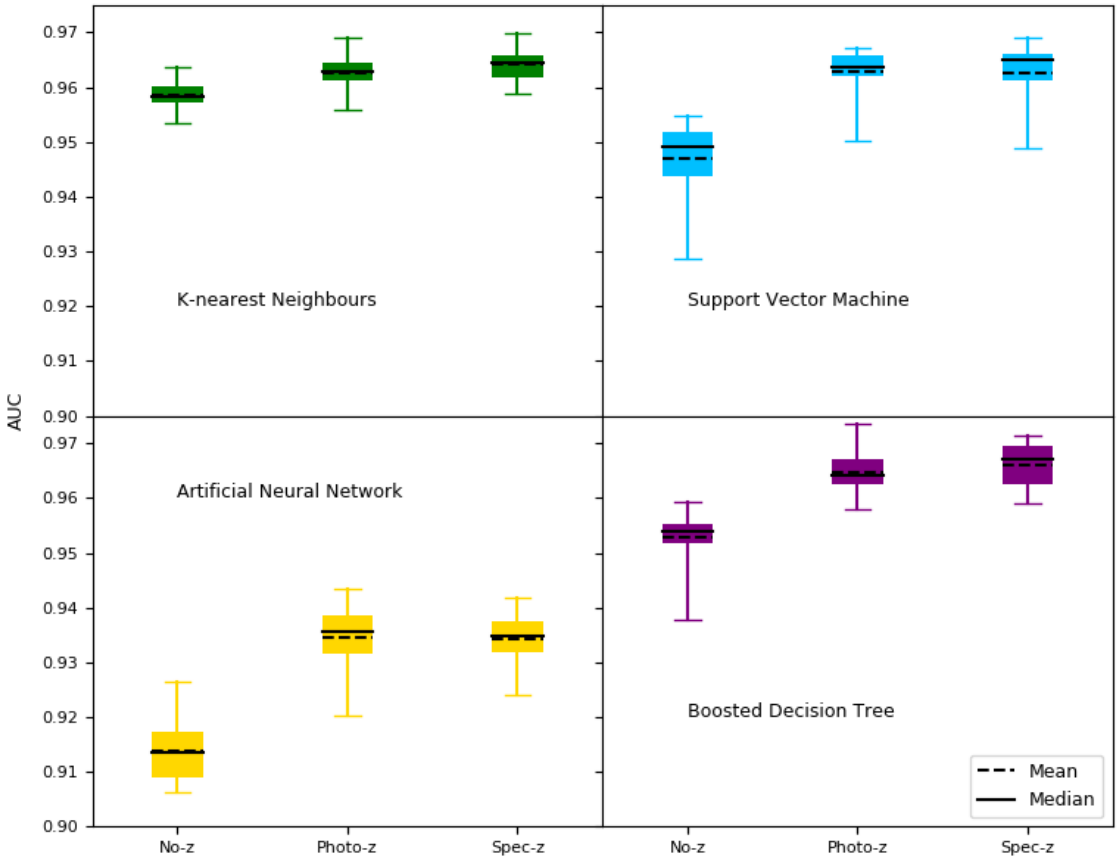


Figure 3.5: Boxplots showing the AUC scores over 20 classification runs for representative training samples comparing the use of no redshift information (No-z), photometric redshifts (Photo-z) and spectroscopic redshifts (Spec-z). The boxes span the interquartile range with whiskers extending out to the full range of AUC values.

Table 3.1: AUC means, medians, interquartile ranges, maxima and minima for representative training samples over 20 runs, and the number of those runs that reached 95% purity. These summarise the results shown in Fig. 3.5 for the four different algorithms, comparing the cases of no redshift (No-z), photometric redshift (Photo-z) and spectroscopic redshift (Spec-z).

Algorithm	Redshift	Mean	Median	IQR	Max	Min	Purity 95%
KNN	No-z	0.959	0.958	0.003	0.964	0.953	20
	Photo-z	0.962	0.963	0.003	0.969	0.956	20
	Spec-z	0.964	0.965	0.004	0.970	0.959	20
SVM	No-z	0.947	0.949	0.008	0.955	0.929	18
	Photo-z	0.963	0.964	0.004	0.967	0.950	11
	Spec-z	0.963	0.965	0.005	0.969	0.949	12
ANN	No-z	0.914	0.913	0.008	0.926	0.906	0
	Photo-z	0.934	0.936	0.007	0.943	0.920	1
	Spec-z	0.934	0.935	0.006	0.942	0.924	1
BDT	No-z	0.953	0.954	0.004	0.959	0.938	20
	Photo-z	0.965	0.964	0.005	0.973	0.958	20
	Spec-z	0.966	0.967	0.007	0.971	0.959	20

we managed to reach our target purity of 95% in three out of four algorithms. The relatively poor performance of ANN is attributed to the fact that these training samples are small compared to the test set. However, neural networks are known to perform well with large training samples (Goodfellow et al. 2016, Section 1.2.2).

Fig. 3.5 shows that, whilst there is overlap in the spread of AUC scores, the trend for all algorithms is an increase in mean and median, suggesting that redshift is a significantly impactful feature to the outcome of classification performance. The extent of improvement seems to be in agreement with the example of ROC curve results in Lochner et al. (2016) (with the exception of KNN): their AUC scores increase by -0.026, 0.016, 0.016 and 0.010 for KNN, SVM, ANN and BDT respectively. We see an increase in the average AUC scores of 0.003, 0.016, 0.020 and 0.012 (comparing No-z and Photo-z). The slight numerical discrepancies in AUCs may be due to splitting classification probabilities between the types Ia, Ibc and II, rather than just Ia and non-Ia as we have done here. Our finding that there is noticeable improvement when including redshift is in contrast to their conclusion that, when considering relative feature importance, redshift is fairly unimportant to this wavelet feature extraction method.

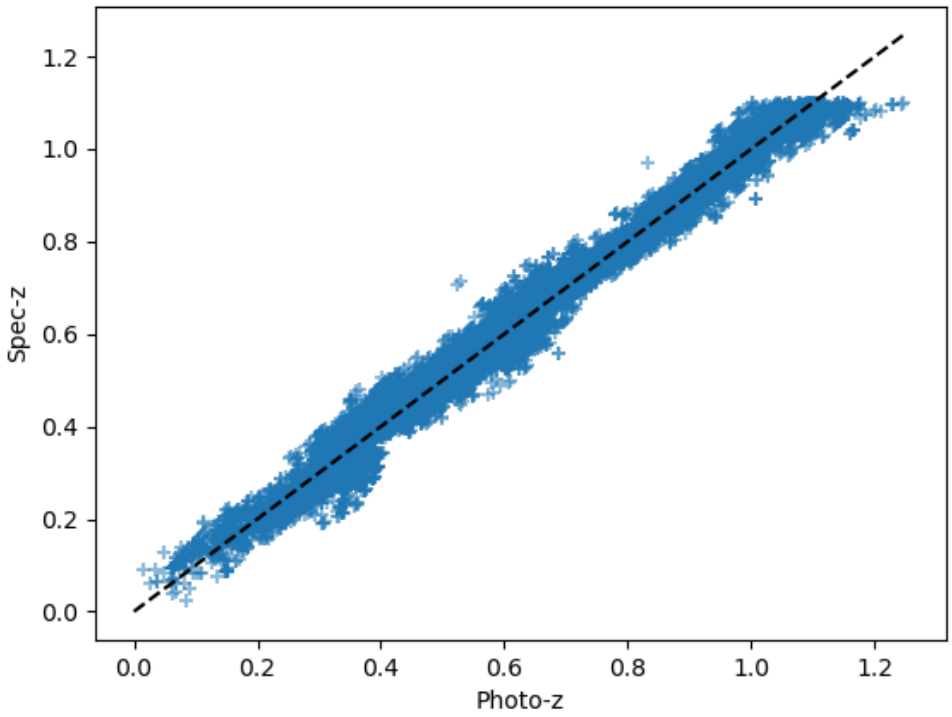


Figure 3.6: Photometric and spectroscopic redshifts of SPCC supernovae. The black dashed line represents equal photometric and spectroscopic redshift values.

We find similar results for photometric and spectroscopic redshift, which may be explained by the absence of any catastrophic outliers in the simulated photometric redshifts in the SPCC; there is little scatter when comparing the two (Fig. 3.6), with a root mean square error of only 0.028. This is perhaps optimistic, as it is estimated that around 10% of galaxy photometric redshift results using LSST photometry will be outliers at $z = 0.5$, reaching even higher percentages at lower redshifts (where outliers are those with redshift error greater than 3 times the robust standard deviation, or 0.06, as defined in Graham et al. 2020).

Furthermore, the study Mitra & Linder (2021) finds that robust supernova cosmology cannot solely rely on photometric redshifts. Investigating the systematic requirements for a LSST-like survey, photometric redshifts at $z \lesssim 0.2$ in particular are found to be problematic, causing bias in dark energy cosmological inference. They conclude photometric redshifts should be used for cosmology only for $z > 0.3$ and spectroscopic classification should be conducted for all supernovae at $z \lesssim 0.2$–0.3.

In this comparison for representative samples we did not alter the photometric redshifts and we used them as they are in the SPCC. Irrespective of the use of either photometric or spectroscopic redshift as an additional feature for classification in this dataset, when the training sample is representative of the test set we observe promising results, including very high purity values. For the rest of this work we use spectroscopic redshifts, as discussed later in § 4.4.3. Our aim is to at least match the classification performance that we would get when using a representative training sample, although, as we show in the next section, current 4MOST capabilities would only deliver a magnitude-limited sample. Our task is consequentially to improve upon a magnitude-limited sample to increase representativity. To address this, we add more training objects at fainter magnitudes and higher redshifts through two routes: observation with larger telescopes and augmentation.

Chapter 4

The Time-Domain Extragalactic Survey

Given the constraints on observing resources for spectroscopic follow-up, we set out to determine how these limited resources would be best used, i.e. how to get the best resulting photometric classification of the remaining sample. In particular we consider the use of the 4MOST (4-metre Multi-Object Spectroscopic Telescope[1]) instrument, which will carry out the Time-Domain Extragalactic Survey (TiDES, Swann et al. 2019), a campaign for spectroscopic follow-up. The follow-up potential with 4MOST is determined by its survey overlap (both angular and temporal) with LSST's observing strategy, its cadence, and TiDES' allocated 250,000 fibre-hours. Despite not being able to follow up every transient event from LSST, TiDES will obtain as many spectra as possible for the purposes of cosmology and creating a basis for a training sample. 4MOST, ESO's newest upcoming spectroscopic facility, is particularly well-suited for this task, with first light expected in 2023. It will be installed on the 4-m VISTA telescope in Chile, at a similar latitude to the Rubin Observatory.

[1] https://www.4most.eu/cms/

4.1 Science goals of TiDES

In the context of this thesis, we particularly consider the TiDES science goals (i) spectroscopic classification of live transients (TiDES-SN) and (ii) spectroscopy of supernova host galaxies (TiDES-Hosts) (Swann et al., 2019).

Science goal (i) enables us to determine which LSST transients are Type Ia supernovae. We can also obtain a spectroscopic redshift from these spectra as the light from the supernovae and their host galaxies will be combined, allowing us to conduct cosmology with these objects. This goal also provides classification of transients (both Ia and non-Ia), which is important for a training sample.

Science goal (ii) will provide us with spectroscopic redshifts of many host galaxies of supernovae observed by LSST for which live spectroscopy was not possible. Hence, these are the transient objects that will define the test sample, i.e. the supernovae that we want to photometrically classify. We will therefore have a spectroscopic redshift for anything that makes it into our cosmological sample. As we will have spectroscopic redshift information for our training and test samples, we use the spectroscopic redshifts of the SPCC simulated supernovae as an ancillary feature in all following classification simulations. This is the same as the spectroscopic redshift mentioned in § 3.5.

Hence, TiDES' full cosmological sample will consist of LSST Type Ia supernovae identified by spectroscopic classification in science goal (i) and combined with those that have a spectroscopic host-galaxy redshift obtained by science goal (ii) and are identified using photometric classification.

4.2 Synergy with LSST

Thanks to the relative proximity of VISTA to the Rubin Observatory, there is naturally a lot of overlap in the regions of the sky that these two facilities can observe. Ideally, TiDES wants to follow LSST's survey strategy (in terms of field pointings) as much as possible in order to quickly obtain live transient spectra close to maximum light. Near the start of this project, investigation of both 4MOST and LSST survey strategies was done to explore the synergy between the surveys.

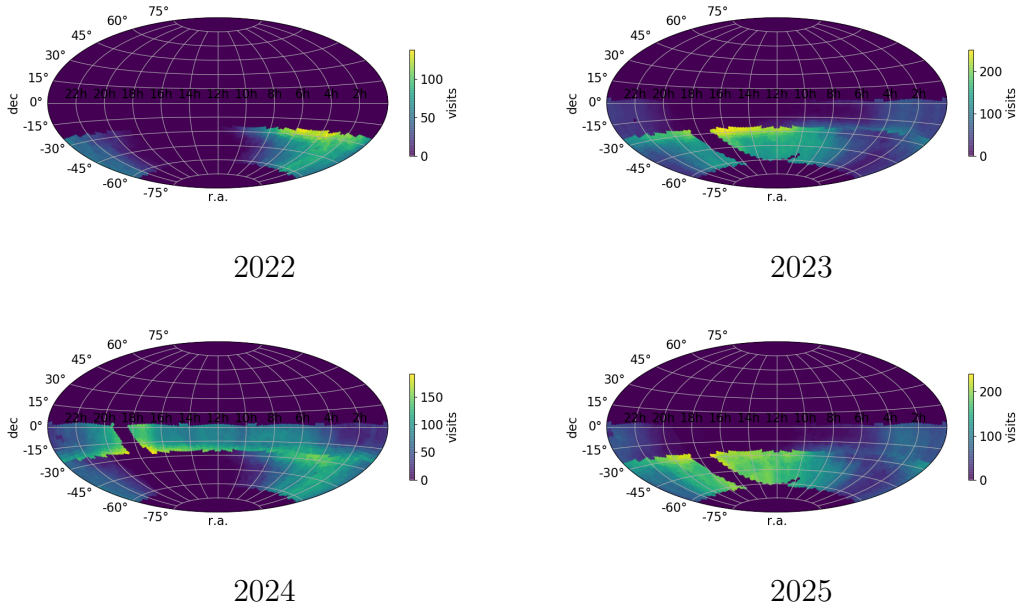


Figure 4.1: LSM for TiDES-SN: maps for observation of live transients. The sky maps are updated for each year to reflect the planned footprint during that year. As LSST begins part-way through the second year of 4MOST's survey, the 2022 map does not have as many observations as the other years.

In order to develop the 4MOST survey strategy, its consortium surveys are required to provide a figure of merit comprising large and small scale merits. The small scale merit (SSM) is based on the survey's ability to meet its science goals. For TiDES, this depends on the success of observing supernovae, host galaxies and AGN. The large scale merit (LSM) is defined as the area of the sky we want to observe. TiDES' LSM is therefore given by the LSST footprint, weighted by the number of LSST observations in a given area. Unlike other 4MOST surveys, TIDES-SN is time-critical and so its LSM is a function of time. As long as host galaxies are identified, they can be observed with 4MOST at any point after the transient event to obtain spectroscopic redshifts. The LSM for host galaxies is therefore the whole WFD survey. For AGN, the LSM is given by the area of LSST DDFs.

Sky maps were created (Figs. 4.1, 4.2 and 4.3) showing the TiDES LSM based

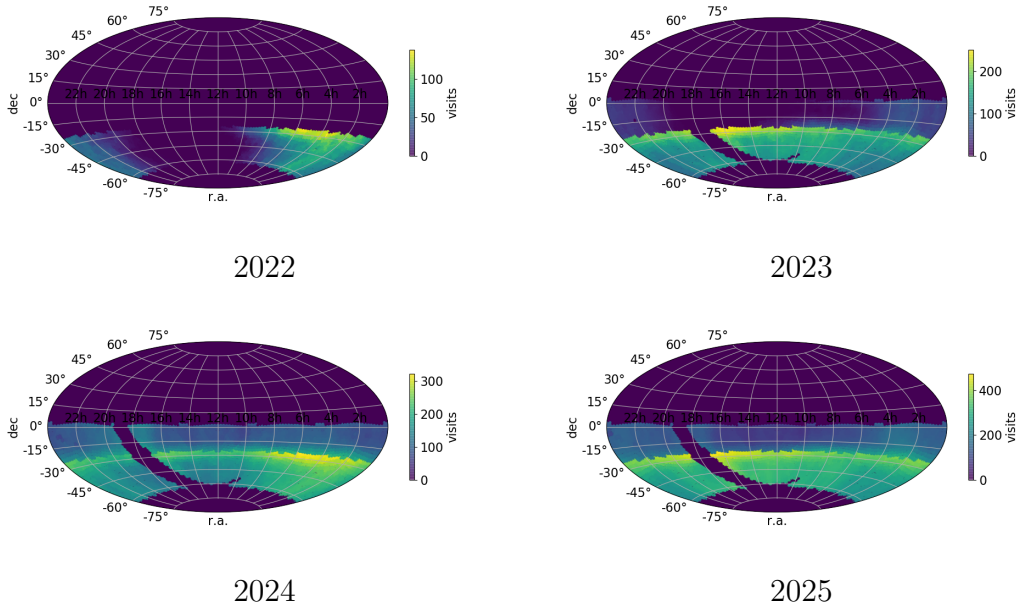


Figure 4.2: LSM for TiDES-Hosts: it does not matter when host galaxies are observed, so LSM is accumulative over each year. The more observations in a field, the more likely we will want to return to obtain spectroscopic redshifts.

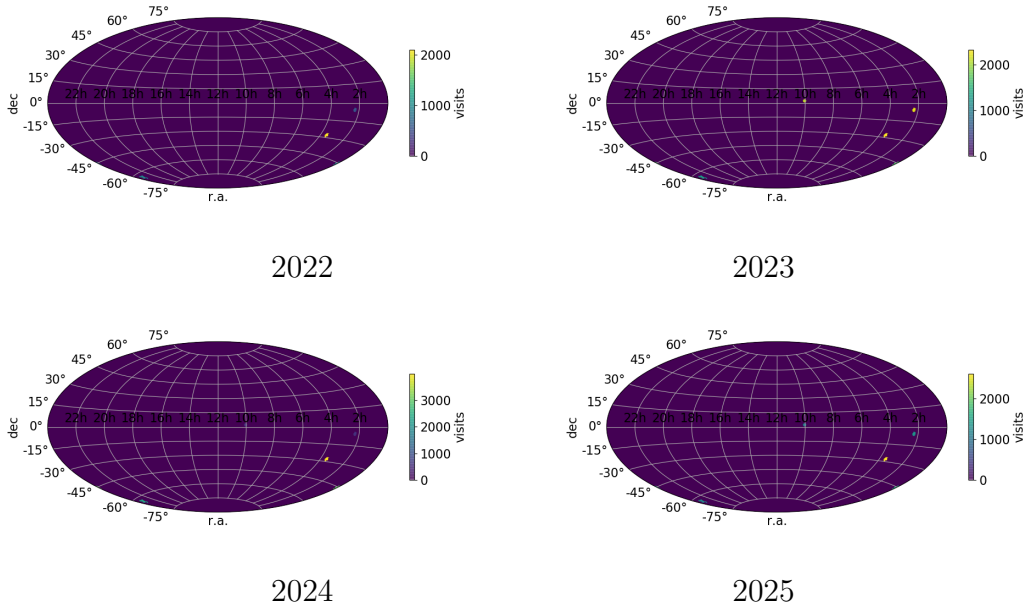


Figure 4.3: LSM for TiDES-RM: AGN reverberation mapping requires observations in LSST's DDFs, a few individual fields that have many visits.

on the LSST `mothra_2045` OpSim simulated survey strategy[1]. This is an example of a rolling cadence strategy, in which LSST's WFD is split into two strips that are alternated each observing year. Rolling cadence enhances sampling frequency by at least a factor of two (LSST Science Collaboration et al., 2017), highly beneficial for obtaining multi-band observations of supernovae light curves. The number of expected supernovae is linearly dependent on both area and observation time (see next section), so covering half of the survey area while doubling field visits will not necessarily result in fewer discoveries. `mothra_2045` was chosen as it fulfils these desired criteria. Note that these simulations are now out of date due to delays with both 4MOST and LSST; at the time when these plots were made 4MOST's simulated survey started at the beginning of 2021 and LSST's started towards the end of 2022. This means the first year and a half of 4MOST had no overlap with LSST. The LSM plots still represent the overall survey area and typical number of field visits and were created for each 4MOST survey year in which there is overlap with LSST, four for each of TiDES' science goals. LSST continues to develop its survey strategy (LSST Science Collaboration et al. 2017 discusses general science-driven optimisation and the papers Lochner et al. 2018; Scolnic et al. 2018a focus on optimisation for dark energy science) and these types of LSM maps are being used to coordinate with 4MOST survey planning. This work is now being continued by others (Frohmaier et al., in preparation).

4.3 Supernova rates

The number of supernovae that we observe in the night sky is limited by several factors. There are observational constraints such as the angular area surveyed, the length of observation time and limiting magnitude of the chosen telescope, but also limitations such as the probability that a supernova explosion will actually occur. These can be combined into an expression that predicts the numbers of supernovae expected, and is typically an empirical formula based on previous supernova surveys. Perrett et al. (2012) determines a volumetric supernova Type Ia rate based on discoveries from the Supernova Legacy Survey (SNLS). The

[1] https://www.lsst.org/scientists/simulations/opsim

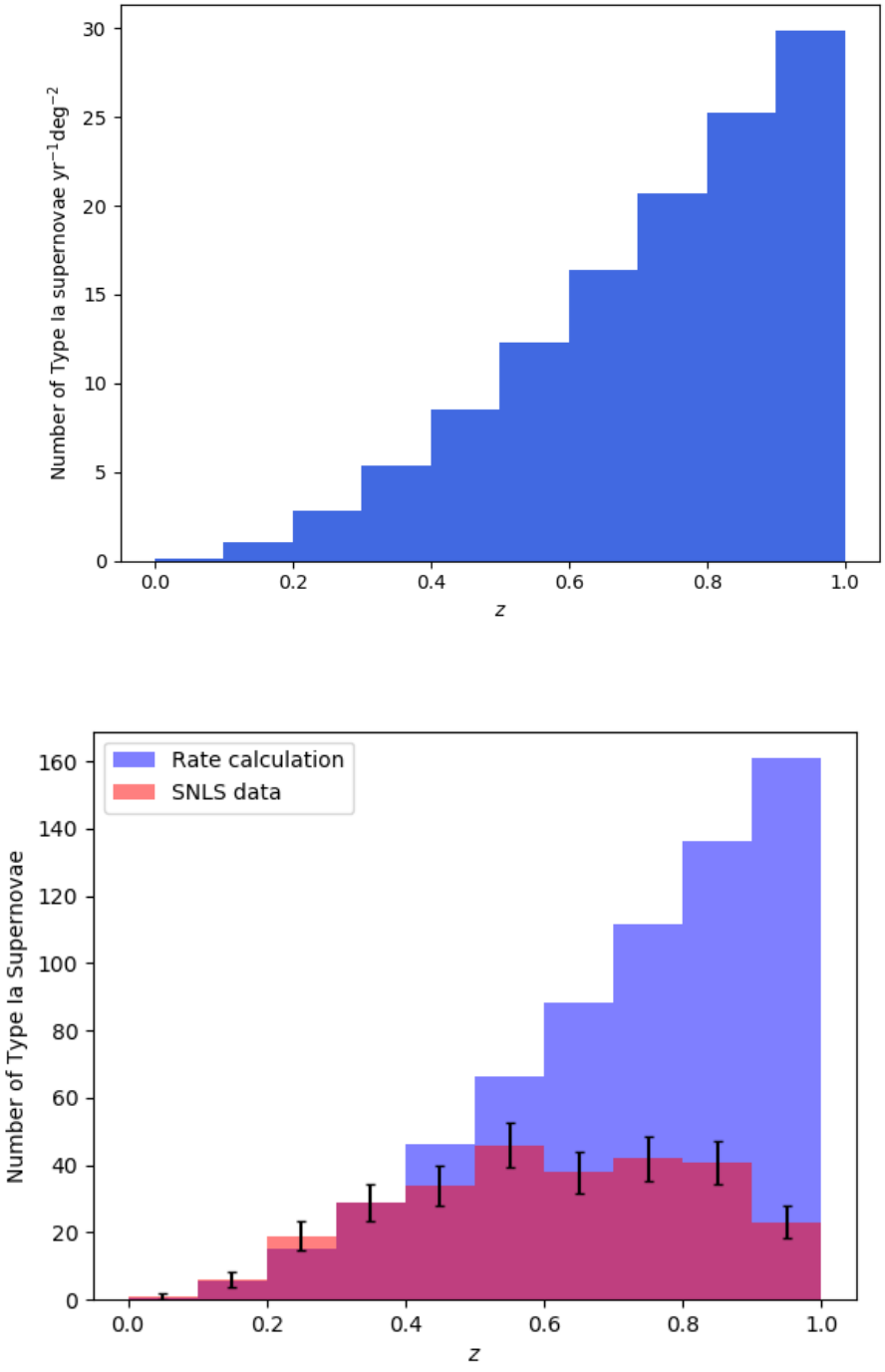


Figure 4.4: Top: Type Ia supernova per deg^2 per year as a function of redshift. Bottom: Type Ia supernova rate calculated based on data from the SNLS. These begin to deviate around $z \approx 0.5$ due to a decreasing detection efficiency. The areas and times used to create this figure are given in Table 4.1. Error bars represent Poisson uncertainty.

Table 4.1: SNLS field areas and times

Field	Area (deg^2)	Time (years)
D1	0.8822	1.2677
D2	0.9005	1.4456
D3	0.8946	1.8400
D4	0.8802	1.5058

rate is volumetric as it considers supernovae over an angular area Θ (in deg^2) and between different redshifts, i.e. the number of supernovae in a redshift bin between z_1 and z_2. Assuming a flat ΛCDM model of the Universe, the volume is calculated as

$$V = \frac{4\pi}{3}\frac{\Theta}{41253}\left[\frac{c}{H_0}\int_{z_1}^{z_2}\frac{dz'}{\sqrt{\Omega_m(1+z')^3+\Omega_\Lambda}}\right]^3 \mathrm{Mpc}^3, \tag{4.1}$$

where c is the speed of light, H_0 is the Hubble constant, Ω_m and Ω_Λ are the density parameters of matter and a cosmological constant. For flat cosmology, Ω_Λ can be replaced by $1-\Omega_m$. In this volume, the expected number of supernovae within time t and at redshift z (taken as $[z_1+z_2]/2$) is given by

$$N = t\epsilon V r_0 (1+z)^{\alpha-1}, \tag{4.2}$$

where r_0 and α are power law parameters determined empirically in Perrett et al. (2012) and ϵ is the detection efficiency. Using this rate calculation, we show the expected number of Type Ia supernovae per deg^2 per year as a function of redshift in Fig. 4.4 (Top).

Going to higher redshifts, the volume in a given area increases and hence includes more host galaxies (until eventually getting to the early Universe), although the detection efficiency decreases as supernovae appear fainter. There is therefore a peak in numbers of discovered supernovae around $z \approx 0.5$ that then drops with increasing redshift and begins to significantly deviate from the actual number of supernovae (the number of observed supernovae is given in Guy et al. 2010). For each SNLS field season, the rate is calculated using the unmasked areas quoted in Perrett et al. (2012) and time is approximated from the range

of observations in Guy et al. (2010). This is summarised in Table 4.1 and the SNLS Type Ia rate is illustrated in Fig. 4.4 (Bottom). We see that the rate calculation and data are in good agreement, confirming the rate calculation, until the reduction in detection efficiency with redshift becomes significant. While the actual number of supernovae only increases with redshift, the proportion detected becomes smaller and smaller after $z \approx 0.5$. The numbers of objects that LSST, and therefore TiDES, expects to find (as quoted in the next section) are calculated in a similar way.

4.3.1 Follow-up strategy

Once 4MOST's survey strategy is finalised, TiDES will need to decide how best to distribute its allocated 250,000 fibre-hours of spectroscopy. TiDES will be exploiting the fact that wherever 4MOST points in the extragalactic sky, there will be LSST live transients to follow up. Hence, rather than driving the 4MOST pointings, TiDES will be 'piggy-backing' on the other surveys as the target density of transients is not high enough for efficient observations on its own. Once receiving LSST transient alerts/detections, TiDES will aim for a rapid turnaround time of 3–4 days in which to target the allocated fibres (approximately 2% of 4MOST fibres) on to these objects and obtain their spectra.

We estimate that TiDES will be able to classify transient spectra to magnitudes as faint as $r_{\rm AB} \approx 22.5$ mag. We explain the origin and implications of this magnitude limit in § 4.4. It will be the main factor influencing the training sample of supernovae we expect to produce using 4MOST. LSST is expected to detect transients fainter than this, making point-source detections down to a depth of $r_{\rm AB} \approx 24$ mag in a single field visit. Consequently, the performance of our classification algorithms depends on how we deal with this magnitude limit.

TiDES will target all live transients ($r_{\rm AB} < 22.5$ mag) in each 4MOST pointing during grey and dark time. Depending on the nature of the final LSST survey strategy, we expect a density of 6–12 live transients per pointing. Over the 5-year duration of TiDES this equates to an expected total of >30,000 transients, with the remaining fibre-hours used to measure host-galaxy redshifts of LSST transients. Final numbers are highly dependent on both LSST and 4MOST survey strategies

that at the time of writing are not yet finalised. TiDES' spectroscopic sample can be used for training our machine-learning algorithms to subsequently classify other LSST transients. The supernova light curves that we will photometrically classify are those for which we have secured host-galaxy redshifts. Combining the Type Ia supernovae in the spectroscopic and photometrically classified samples, altogether, TiDES therefore expects to produce the largest cosmological sample of Type Ia supernovae by over an order of magnitude.

Classifying live supernovae that are fainter than 4MOST's limit would require use of 8-m and larger telescope facilities, such as the Very Large Telescope[1] (VLT) and the upcoming, next generation Extremely Large Telescope[2] (ELT), Thirty Meter Telescope[3] (TMT) and Giant Magellan Telescope[4] (GMT). However, to classify live supernovae, time on these telescopes is likely to be even more limited than on 4MOST, so we do not expect more than a few hundred sources to be observed. We return to this in § 5.2.1.

4.4 Simulating a 4MOST spectroscopic sample

We set out to simulate a TiDES spectroscopic training sample to use with our chosen classification software SNMACHINE using the SPCC dataset.

We assume that 4MOST will have 1 h field visits based on Tempel et al. (2020a). For TiDES, a field visit exposure time of 1 h in combination with a spectral success criterion (SSC) effectively imposes a magnitude limit to spectroscopically confirmed supernovae. Here we discuss how to simulate our spectroscopic sample by using the 4MOST capabilities as a guide.

4.4.1 TiDES simulations

To determine the limiting magnitude for 1 h exposures, we use a realistic mock catalogue containing supernovae with population fractions following the literature.

[1] https://www.eso.org/public/teles-instr/paranal-observatory/vlt/
[2] https://www.eso.org/public/teles-instr/elt/
[3] https://www.tmt.org/
[4] https://www.gmto.org/

Included supernova types are: Ia, split into normal Ia, 91T and 91bg using the fractions of each type given in Li et al. (2011) and with a rate from Frohmaier et al. (2019); Core-collapse, split into Ib, Ic, IIL and IIP using the fractions given in Richardson et al. (2014), with a rate proportional to the star-formation history in Li (2008), anchored at low-redshift by the volumetric core-collapse rate from the Sloan Digital Sky Survey II Supernova Survey (Taylor et al., 2014). The different supernova types and rates in the catalogue are necessary to reflect variations in spectra, which affect the rate of success in obtaining spectra of sufficient signal-to-noise ratio (SNR), defined later in this section. The LSST cadence assumed follows the `mothra_2045` OpSim survey strategy. The catalogue itself is limited at a peak magnitude of $r_{\mathrm{AB}} = 24$ mag, where any supernova that peaks brighter than this is simulated to be detected by LSST.

Each transient in the catalogue is assigned a spectrum from a set of templates based on its type, phase and magnitude. Additionally, for Type Ia supernovae, there is variation in their spectra based on the x_0, x_1 and c SALT2 light curve parameters (Guy et al., 2007). The spectra, normalised to the r-band magnitude at the time of observation, are run through the 4MOST exposure time calculator (ETC), which can quickly calculate exposure time requirements for thousands of targets for a given SSC. The ETC uses the 4MOST instrument response and outputs of the simulator TOAD (Top-Of-Atmosphere-to-Detector; Winkler et al. 2014), providing extensive modelling of both system throughput and sensitivity. The ETC is a parametrised version of TOAD, calculating the 1D signal and noisy spectra for targets with different target-fibre alignments and observing conditions such as sky brightness, transmission and seeing. By specifying a SNR (and given the magnitude of the targeted supernova), the ETC can return the target's required exposure time (and vice versa). We later use this to calculate the fraction of sources successfully observed within our assumed 1 h exposure time.

4.4.2 Spectral success criterion

Our results come from running the catalogue through the 4MOST ETC v0.02 (in May 2019). However, since then, the ETC has been updated with newer versions. For a fixed exposure time and scaling results to the same effective SNR criterion,

Table 4.2: The model for spectral success with 4MOST used to define the probability that an object observed with magnitude $r_{\rm AB}$ will be selected for our simulated training sample.

Magnitude	Success rate
$r_{\rm AB} < 21.75$	1.0
$21.75 \leq r_{\rm AB} \leq 22.75$	$\{1 + \exp[10\ (r_{\rm AB} - 22.25)]\}^{-1}$
$r_{\rm AB} > 22.75$	0.0

we find that the ETC v0.6 (in September 2020) agrees with the ETC v0.02 to within 0.02 mag, and so the difference was ignored. For TiDES supernovae, given a SSC, the success of observation depends upon both the spectral features present and the amount of 'contaminating' light from the transient host galaxy (Swann et al., 2019). As supernova spectra are dominated by broad features, the TiDES SSC is defined using the average SNR over 15 Å bins (over the range 4,500-8,000 Å in the observed frame). The TiDES criterion is based on earlier studies of high-redshift Type Ia supernovae (Balland et al., 2009), where robust classification is achieved with a mean SNR = 5 per 15 Å and probable classification of transients is demonstrated with a mean SNR as low as 3 per 15 Å. However, in this study we adopt a more conservative criterion of SNR = 7 per 15 Å. We assume that all spectra that meet this criterion are correctly classified in our tests of optimising photometric classification using the TiDES sample (Chapter 5). Following these tests, we investigate contamination of the spectroscopic sample, discussed in Chapter 6.

Current 4MOST simulations combine observations of the same sky coordinates and instrument position angle into observing blocks (OBs, Tempel et al. 2020a). The duration of the OBs are limited by a total exposure time of 1 h. Success is determined by whether a targeted supernova spectrum's necessary exposure time falls below 1 h for our criterion of SNR = 7 per 15 Å. The rate of success for obtaining supernova classification from their spectra as a function of magnitude is shown in Fig. 4.5. The success rate does not take into account 4MOST's fibre-target allocation which will depend on all 4MOST surveys and their science goals (Tempel et al., 2020b). Observation of each object in the catalogue was simulated in dark and grey time (we assume none of our targets will be targeted during

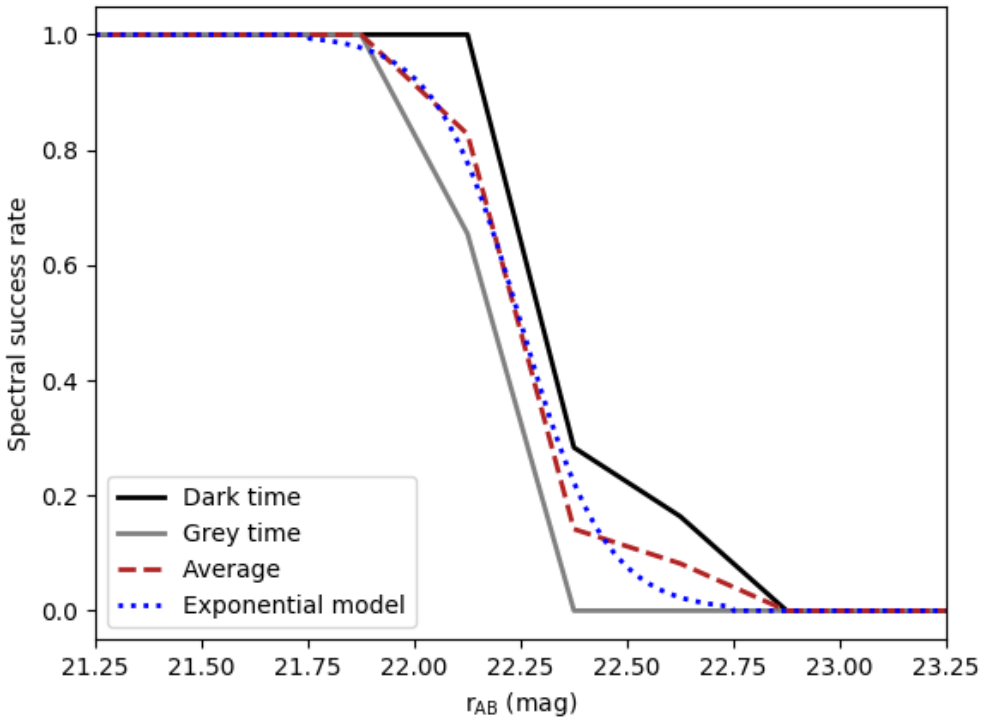


Figure 4.5: The success rate represents the probability that we obtain a spectrum of sufficient signal to noise, and therefore successful classification, of a targeted supernova of magnitude r_{AB}. The success rate is 50% at $r_{\mathrm{AB}} = 22.25$ mag. The rate is calculated as the proportion of input supernovae for which a successful spectrum was obtained in magnitude bins of size 0.25. The average between the rates for dark and grey time is modelled as an exponential function (see Table 4.2) that is later used to create our simulated training sample.

bright time). Dark and grey time are defined by the amount of moon illumination (fraction of lunar illumination, FLI) where FLI < 0.4 and $0.4 \leq$ FLI ≤ 0.7 for dark and grey respectively[1]. The success rate is averaged over both curve distributions at each magnitude as the dark/grey/bright is undecided for 4MOST. However, current simulations for 4MOST's tiling pattern favour dark time over grey (Tempel et al., 2020a), so their average can be considered as a lower limit to our success rate. The function describing the success rate is shown in Table 4.2. The exponential function in the second row was chosen to represent the average between dark- and grey-time success rates.

With 4MOST, it may be that we do not get the full 1 h observation for all our supernovae. This exposure time is based on two 30 min exposures in a single OB. Splitting into single exposures will affect the success rate of obtaining spectra of live transients. For the extreme (and unlikely) case in which all OBs contain only single 30 min exposures, the success rate curve keeps the same shape but moves ∼0.5 mag towards brighter magnitudes, i.e. 50% success rate occurs at $r_{\mathrm{AB}} = 21.75$ mag. This would be much less favourable for our training sample prospects than for the success rate we simulate in Fig. 4.5.

4.4.3 Use of redshift

In general, 4MOST will not give us the opportunity to return to the same pointing of previously observed live transients and obtain a pure host-galaxy redshift. However, when we observe live transients, the light from the supernova and host galaxy will be blended and we expect to be able to measure host redshifts from these spectra, although not necessarily other host properties. This is what allows us to use the Type Ia in our spectroscopic sample for cosmology.

As we will have spectroscopic redshift information for our training and test samples, we used the spectroscopic redshifts of the SPCC simulated supernovae as an ancillary feature. This is the same as the spectroscopic redshift mentioned in Section 3.5 and is used throughout all following photometric classification simulations unless otherwise stated.

[1] https://www.eso.org/sci/observing/phase2/ObsConditions.html

4.4.4 Creating the training sample

Having determined the magnitude limit to a TiDES spectroscopic sample of supernovae, we apply this to the SPCC dataset to create a simulated training sample to be used in SNMACHINE. SPCC supernovae are selected for inclusion in the simulated training sample with a probability that follows the model described in Table 4.2. However, to avoid using all supernovae brighter than the magnitude limit in the training and therefore not leaving any bright objects in the test set, the probabilities are scaled down by a factor of 2. This results in a magnitude-limited training sample of approximately 500 supernovae. Given that we are expecting a spectroscopic sample of size >30,000 from TiDES, we would require a much larger dataset to fully simulate our prospective results. Nevertheless, by applying the 4MOST magnitude limit we are investigating its effect on algorithm-training and, ultimately, how to maximise our classification potential based on this observing constraint.

When selecting the training sample based on magnitude, the magnitudes used for each supernova are the r_{AB} magnitudes closest to peak (i.e. the brightest r-band observation in the simulated light curve). By making the connection with Fig. 4.5, which uses magnitude at the time of observation, we are assuming that we will obtain supernova spectra close to peak, within a few days. This is acceptable as, given 4MOST's limiting magnitude, we can only hope to get most objects' spectra close to peak.

Fig. 4.6 shows a stacked magnitude histogram of the training and test sets for one such magnitude-limited example. These training samples are referred to as MagLim to distinguish from others later on. We also show the distribution of training objects with respect to the test set (remaining objects from the SPCC) in redshift-magnitude space (Fig. 4.7). Comparing to the representative training sample example in § 3.4 (Figs. 3.1 and 3.2), clearly, a magnitude-limited training sample is not covering the full ranges of redshift and magnitude present in the test set. We examine the effect this has on the feature space of the training supernovae with respect to the test set in § 5.3. We run this same test for 10 different training sets, sampled in the same way.

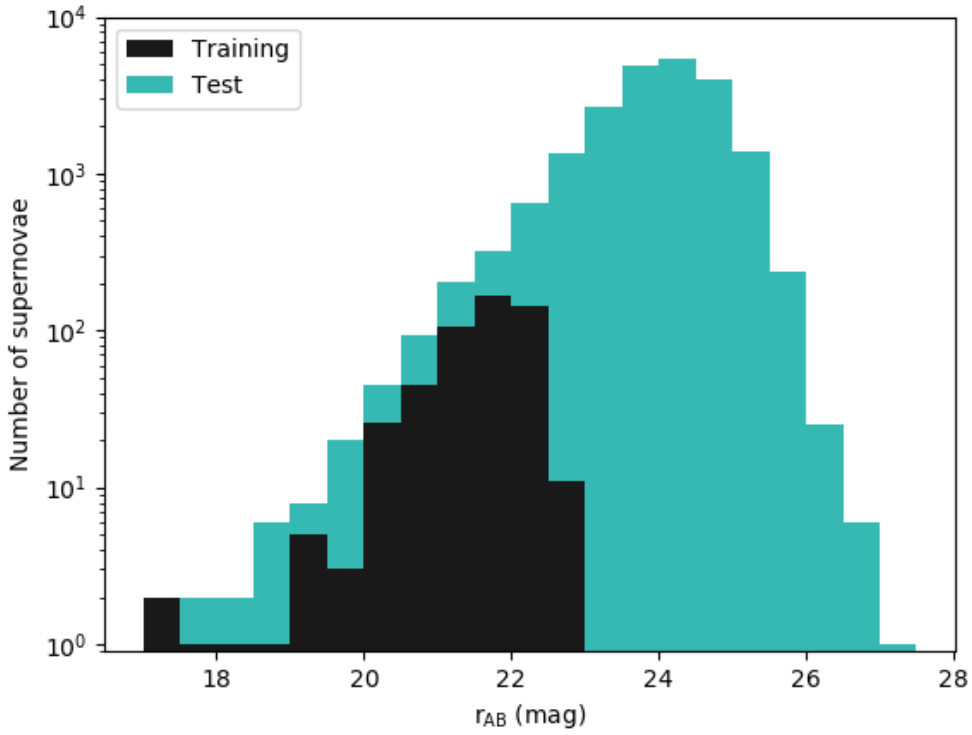


Figure 4.6: Stacked magnitude histogram of a magnitude-limited training sample (MagLim) and test set.

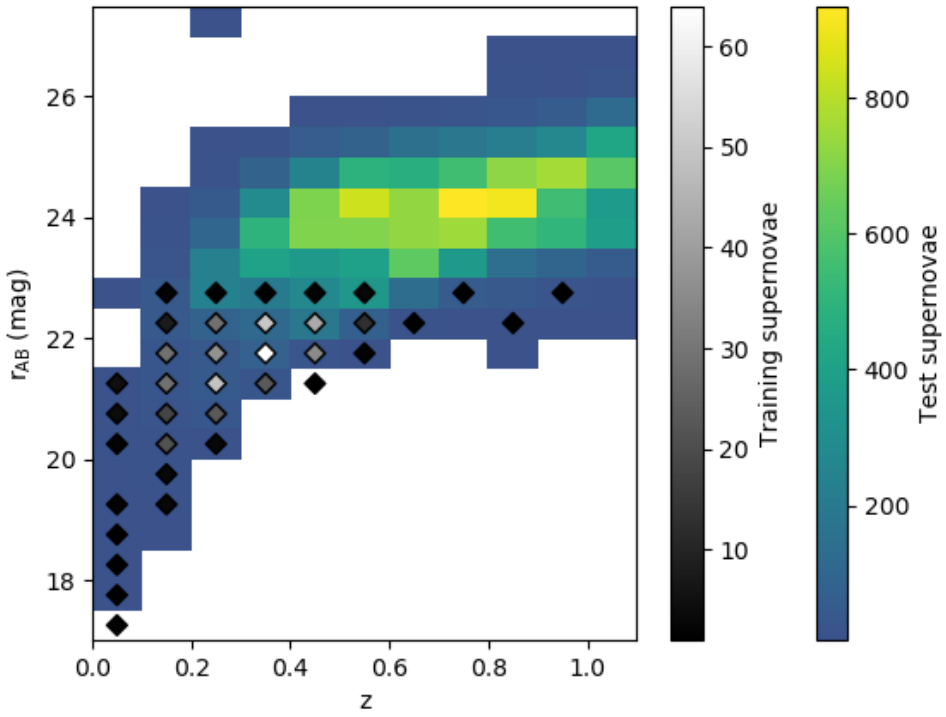


Figure 4.7: 2D histogram of the relative distributions of redshift and magnitude in magnitude-limited training (MagLim) and test set.

Chapter 5

Optimising the TiDES Training Sample

This chapter is dedicated to presentation of the main results from our classification simulations, starting with the 4MOST magnitude-limited training sample. All our results are summarised in Table 5.1 at the end of the chapter.

5.1 Results of magnitude-limited training

We repeat the classification process using the SNMACHINE pipeline as discussed in § 3, now using the TiDES magnitude-limited training sample from § 4.4.4 instead of a representative one.

Considering the ROC curves for this training sample (Fig. 5.1, left), we find that the classifiers struggle to perform much better than random (shown by the dashed line). A good classifier that achieves high TPR and low FPR would be as far from the dashed line as possible, reaching the top-left corner. For KNN, SVM, ANN and BDT, the average AUC scores are 0.547, 0.671, 0.702 and 0.628 respectively and the results are summarised in Table 5.1 in the MagLim row. The magnitude limit has evidently had a negative impact on the classification.

We find that it is also difficult to reach high purities (Fig. 5.1, right). Often, it is impossible to reach a purity of 95% and, even the few times it succeeded

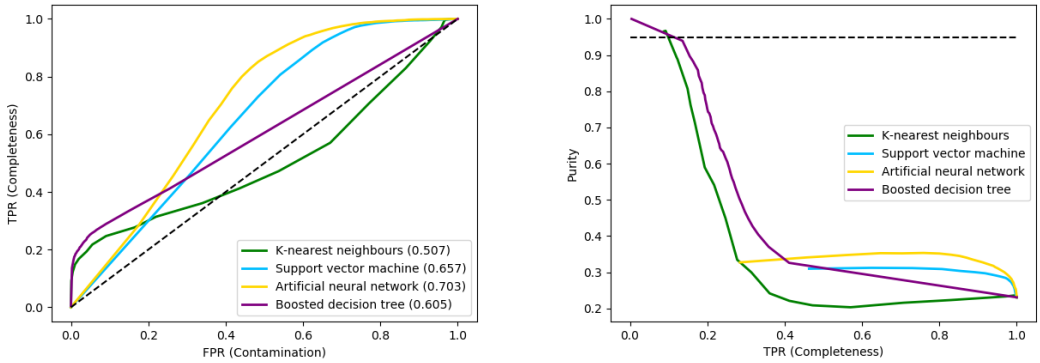


Figure 5.1: For a magnitude-limited training sample (MagLim), ROC curves are close to resembling those of random classification. Despite the high purity reached for KNN (just reaching the 95% target) and BDT, the returned completeness of the classified sample is very low. Comparing to representative training, we are far from the classification algorithms' potential and need to improve upon this training sample.

(generally requiring the maximum probability threshold), we return so few correct Type Ia supernovae that manipulating the classification parameters to achieve this would not significantly increase our cosmological sample. Also, not shown in the figure, we find that the completeness for high purities is consistently zero much beyond the faintest magnitudes of the training supernovae. For more practical uses, we instead therefore require methods to address this bias towards bright, low-redshift supernovae and produce a more representative training sample.

The bias towards bright objects if we use only 4MOST implies a non-representative training sample. TiDES is currently planning to blanket target every transient possible that is brighter than $r_{\mathrm{AB}} = 22.5$ mag. However, this is a form of Malmquist bias, as the spectroscopic sample's preference towards brighter objects is a result of observational capabilities; it is much easier to obtain spectra of bright supernovae as it takes much less time to reach the desired SNR (we saw this previously in our discussion of supernova rates, as magnitude and redshift are strongly correlated for Type Ia supernovae and therefore detection efficiency decreases with redshift – Fig. 4.4).

It is interesting that the algorithms perform so poorly with wavelet features. Wavelet transforms are approximately invariant under translation and stretch,

suggesting that similar supernovae will have similar features, even with different explosion times and redshifts (Lochner et al., 2016). It appears that, despite this, the magnitude-limited sample does not have enough information from its objects' features to be able to classify fainter supernovae. We analyse the feature space coverage of different training samples in § 5.3.

5.1.1 Redshift in magnitude-limited training

It was previously discussed that spectroscopic redshift is required for cosmology. However, this does not necessarily mean that we need to include it in the classification step. While we find that including either a photometric or spectroscopic redshift will improve classification results when used on representative training samples (§ 3.5), it may not be the case when our training sample is non-representative.

The magnitude limit seems to also imply a redshift limit (very few, if any, training supernovae are found beyond $z = 0.5$–0.6, shown in Fig. 4.7), although, depending on the specific sampling, the cut-off may not be as obvious. Including redshift in a magnitude-limited training sample does not give any extra information about fainter supernovae, and may be introducing further bias to the sample as it only adds extra information that is also non-representative of the test set. On the other hand, redshift information may improve classification accuracy at brighter magnitudes. We therefore investigate the effects of withholding redshift information in our previously simulated magnitude-limited training samples.

Inclusion of spectroscopic redshift, as opposed to none, in magnitude-limited training samples does not make a clear improvement to classification (comparing MagLim and MagLimNo-z in Table 5.1), as it did for representative training. Without redshift information, SVM and ANN perform worse and KNN and BDT seem to improve based on their mean and median scores. All four algorithms have a wide range of results, although they yield a higher maximum AUC when redshift is included.

5.2 Reaching fainter magnitudes: training beyond 4MOST's limit

5.2.1 Making use of larger spectroscopic facilities

4MOST alone cannot provide us with a fully representative training sample. The required exposure times for supernovae fainter than $r_{\rm AB} \approx 22.5$ mag are generally too large to consider spectroscopic follow-up with 4MOST. One option would be to use other spectroscopic facilities such as the VLT and ELT.

We first explored the possibility that we may only need a few fainter objects to combine with our 4MOST sample. As a machine-learning exercise, we investigated the effect of adding one randomly selected supernovae of Type Ia, Ibc and II in each 0.5 magnitude bin above the 22.5 mag limit (to the faintest magnitude bin containing only one Type II supernova). This provides an additional 28 faint supernovae. Assuming using the VLT up to $r_{\rm AB} = 24$ mag and the ELT beyond this, we estimate the total required exposure times of approximately 80 h and 90 h respectively, for all 28 of these specific objects. For comparison, a typical magnitude-limited training sample in the SPCC ($\sim$500 objects) would require a total exposure time of $\sim$2,200 h with 4MOST. Compared to the original magnitude-limited training sample, over the 10 runs the average AUC increased from 0.554 to 0.760 for KNN, 0.667 to 0.769 for SVM, 0.700 to 0.758 for ANN and 0.623 to 0.769 for BDT, although the algorithms still struggle to reach 95%. By adding just a few faint supernovae into the training, this method shows that a relatively small increase in the number of supernovae can significantly improve upon a magnitude-limited sample. However, this clearly needs to go further and does not compare to our realistic faint sample. Hence, we went on to creating a sophisticated, realistic faint sample of supernovae.

Using the same SNR criterion as for 4MOST, we simulate a more realistic faint spectroscopic sample of supernovae to combine with our 4MOST sample (MagLim+Faint). We simulate a total exposure time of 1,000 h on the VLT and 100 h on the ELT (assuming 100 h and 10 h per 6-month semester over 5 years respectively). Individual supernova exposure times are determined using their

brightest magnitudes and are based on calculations from the online ETCs. For the VLT ETC[1] we use the FORS2 instrument and fixed object (point source at $z = 0.6$, with input flux distribution a power law with index 0; flux $\propto$ wavelength) and sky parameters, varying the magnitude normalised in the r-band to estimate our exposure times. Parameters used are a moon FLI = 0.2, airmass = 1.50, seeing/image quality IQ = 0.80 arcsec with a slit width of 1.00 arcsec using the GRIS_300V+10 (>450nm,GG435) grism. Similarly, for the ELT ETC[2] we use airmass = 1.50 and seeing = 0.8 arcsec with the Laser-Tomography Adaptive Optics mode and radius of circular SNR ref. area = 200 mas.

We randomly sample supernovae from the SPCC between $r_{\rm AB} = 22.5$–24.35 for the VLT and calculate their exposure times (adding an assumed overhead time of 5 min per object) until reaching the total. Supernovae with magnitudes of $r_{\rm AB} > 24.35$ would require > 4 h exposure time with the VLT. Exposures longer than 4 h are possible on the VLT, although it is not practical to do this for many objects in a survey. In our simulation these sources would be observed by the ELT. Hence, we similarly determine a ELT sample of supernovae with $r_{\rm AB} > 24.35$ and 9 min overheads[3]. This produces a sample of ∼600 VLT supernovae and ∼400 ELT supernovae, increasing the total size of our simulated spectroscopic training sample by approximately 200% (∼500 to ∼1500). We acknowledge that in reality our magnitude-limited TiDES sample would be much larger, and hence our simulated TiDES sample is out of proportion to this realistic faint sample. A full accurately simulated sample is not possible with this dataset because of the relatively small number of bright objects.

With the addition of our faint sample we obtain the resulting magnitude and redshift-magnitude distributions in Figs. 5.2 and 5.3. We see a clear improvement on the overall performance of the machine-learning algorithms due to them making more informed classifications; the ROC curves (Fig. 5.4, left) have moved far away from the random classification associated with the diagonal dashed line and

[1] https://www.eso.org/observing/etc/bin/gen/form?INS.NAME=FORS+INS.MODE=spectro

[2] https://www.eso.org/observing/etc/bin/gen/form?INS.NAME=ELT+INS.MODE=swspectr

[3] 6 min for guide star aquisition plus 3 min for the adaptive optics to produce the required image quality, described in the ELT Top Level Requirements at https://www.eso.org/sci/facilities/eelt/docs/index.html

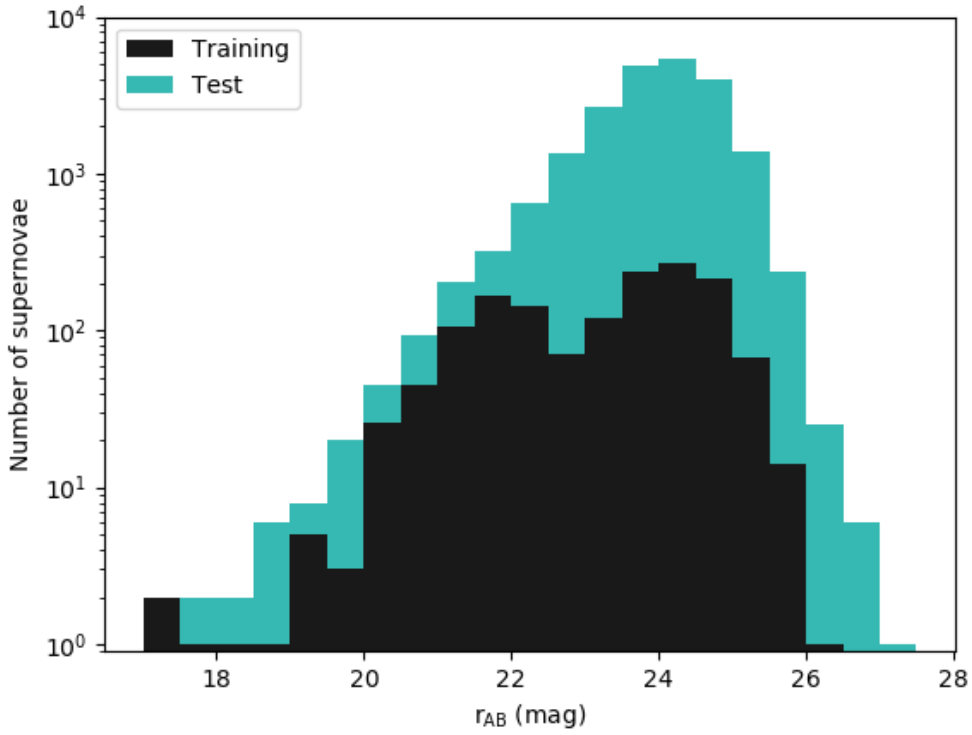


Figure 5.2: Stacked magnitude histogram of a magnitude-limited training sample with an additional faint sample (MagLim+Faint) and test set.

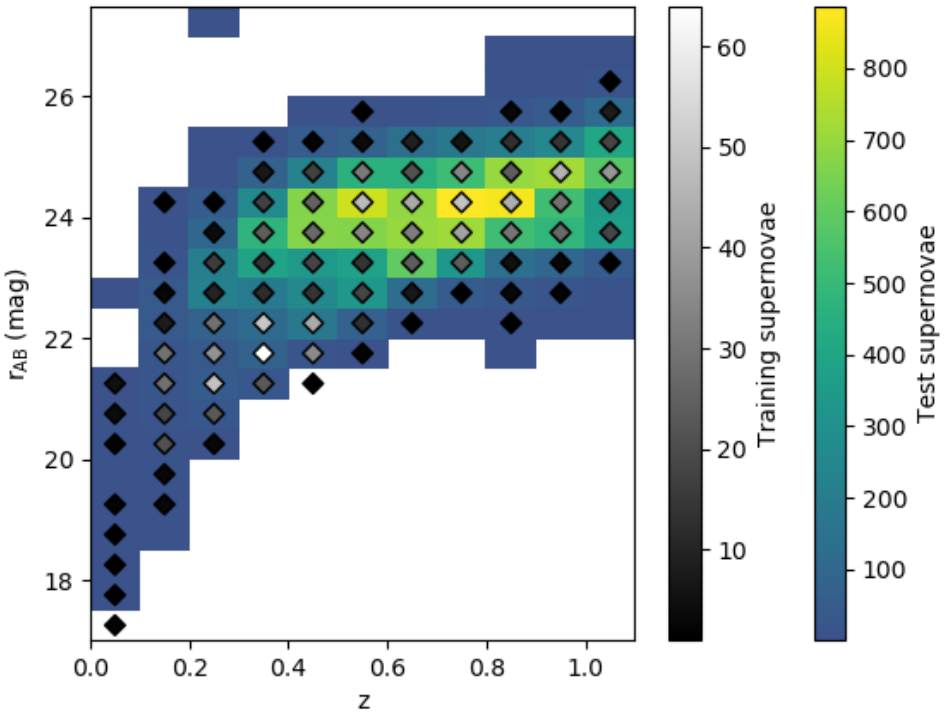


Figure 5.3: 2D histogram of the relative distributions of redshift and magnitude in a magnitude-limited plus faint training sample (MagLim+Faint) and test set.

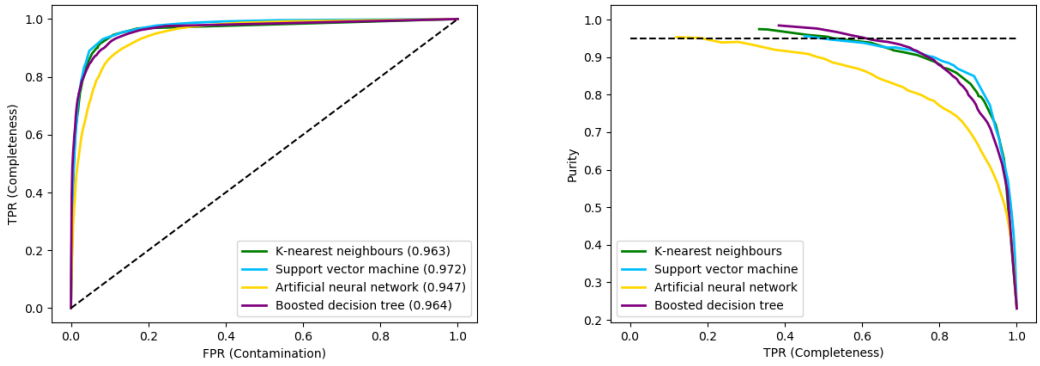


Figure 5.4: ROC and purity curves for a magnitude-limited plus faint training sample (MagLim+Faint). Whilst we still don't have the same distributions of magnitude and redshift as for a representative example, the introduction of a faint sample intro the training has a positive impact on the classification results.

towards the top-left corner. Going from the purely magnitude-limited training to the addition of fainter supernovae, over the 10 runs the average AUC increased from 0.547 to 0.961 for KNN, 0.671 to 0.960 for SVM, 0.702 to 0.946 for ANN and 0.628 to 0.969 for BDT.

Furthermore, we see significant improvements in the purity of our classified samples. Similar to the ROC curves reaching the top-left of the plot, good classification is also indicated by purity-completeness curves reaching the top-right, such as in the example in Fig. 5.4 (right). Notably, adding our faint sample into the training results in all 10 runs reaching 95% purity for KNN and BDT (up from 2 and 7 respectively).

5.2.2 Data augmentation

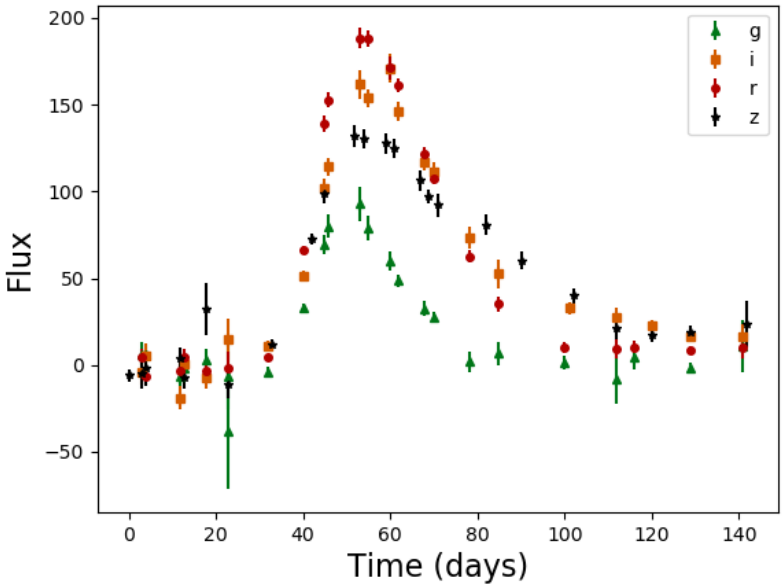
There is another avenue that can be taken to reach fainter magnitudes for our training sample. A fully representative spectroscopic training sample may not be necessary with the advent of data augmentation methods. Revsbech et al. (2018) first demonstrated success in data augmentation of supernova light curves to increase representivity of training samples using their STACCATO model. New training data is generated by drawing from Gaussian processes that are modelled to fit the original light curves.

A similar procedure was applied in the winning solution to PLAsTiCC; Boone (2019) demonstrates with the software AVOCADO that using expensive spectroscopic resources is not required when there are well-sampled, intermediate-redshift objects available for augmenting the training set.

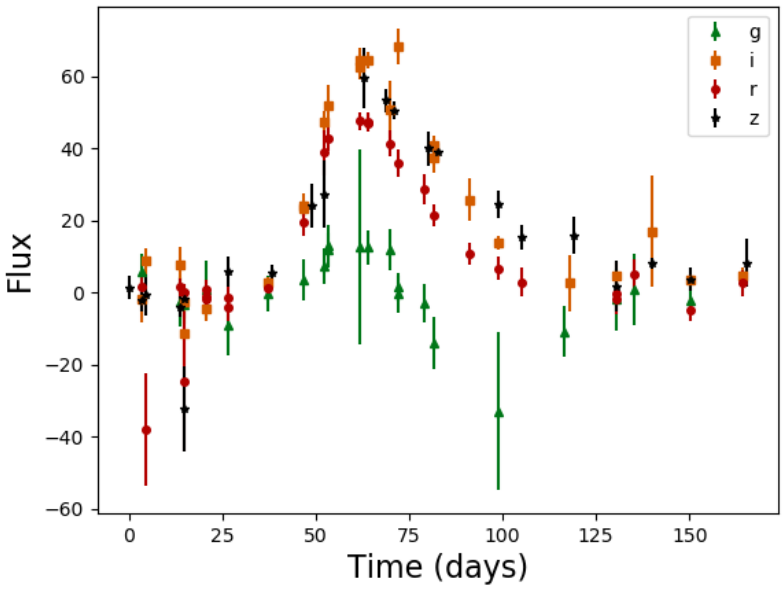
In our case using the SPCC, the test set does not include any classes of objects that are not present in the training sample. If there are previously unforeseen objects in the test set that are not in the training sample, then augmentation cannot help. In the PLAsTiCC competition this was observed with the unknown class 99, which was featured in the challenge's test set but did not exist in the provided training sample (Kessler et al., 2019). Part of the challenge required participants to return classification probabilities for this 'Other' class and successful algorithms made an effort to classify them, although all classifiers struggled to identify these unknown objects (Hložek et al., 2020).

We adapted the source code, AVOCADO, to augment our magnitude-limited training sample by creating new artificial light curves that are resampled, shifted in time, and are at different redshifts for a range of observing conditions and uncertainties (our version is now included in the AVOCADO GitHub)[1]. We use the same augmentation procedure of implementing a 2D Gaussian process (dimensions of time and wavelength), although we make certain changes to AVOCADO, so that our augmented light curves are specific to our dataset and reflect the kinds of light curve that we want to classify. Firstly, we change the band central wavelengths to those of DES to match the SPCC light curves that we are using in the tests. These are used as wavelength coordinates in the Gaussian process. We also ensure that our augmented light curves have a number of observations consistent with SPCC. This is achieved by randomly sampling from a two-peaked distribution used to model the number of light curve observations in the original dataset. We use the same AVOCADO constraints on augmented supernova redshifts to avoid the Gaussian process having to extrapolate far from the available data, where modelling uncertainties dominate its prediction ($0.95z_{\text{old}} < z_{\text{new}} < 5z_{\text{old}}$ and $1 + z_{\text{new}} < 1.5(1 + z_{\text{old}})$, explained fully in Boone 2019). The next part we change is the simulation of the light curve uncertainties. As with the original method in AVOCADO, all the SPCC's error bars in each band are well-modelled as lognormal distributions and so we use the lognormal parameters for our dataset's band noises to sample flux errors and set the depth of observations in our new light curves. Finally, we implemented a method to check whether a new light curve would be useful in the context of our dataset and simulations. The pass criterion is that the new light curve contains simulated observations in the r-band, including a positive maximum flux (used to give the supernova's reference magnitude). Additionally, we discard any augmented light curves that have redshifts and magnitudes that fall outside the ranges in the SPCC. We do not have need of the original AVOCADO methods of preprocessing light curves (accounting for consistent background flux levels) or augmenting galactic objects (objects in the PLAsTiCC dataset that have $z = 0$). Fig. 5.5 shows an example of a supernova light curve with one of its augmented counterparts.

[1] https://github.com/kboone/avocado



$z = 0.45$



$z = 0.69$

Figure 5.5: Top: Original light curve (DES SN 013866), a Type Ia supernova at $z = 0.45$. Bottom: An augmented version of the original light curve, simulated at $z = 0.69$.

For augmenting our magnitude-limited training sample, we use the 2D Gaussian process method in AVOCADO to create up to 50 new versions of each original training supernova. We found that 50 is sufficient by augmenting our magnitude-limited training sample in multiples of 10 from 10 to 100, and seeing that classification (AUC) plateaus for around 40–50 new objects per original light curve. We do not reuse the same augmented light curves, but instead create a new set of augmented light curves for each run.

As augmentation simulates new objects at different redshifts, it therefore requires initial cosmological assumptions[1]. Before using such a method in a real cosmological analysis, it will be important to test (with simulations) the impact of these assumptions on the final cosmological results.

The first augmented training samples we create are from our magnitude-limited samples discussed in § 5.1. For these we augment the training to extend to fainter magnitudes and higher redshifts as shown in Figs. 5.6 and 5.7 (MagLimAug). Without using any of the original SPCC supernovae beyond $r_{\mathrm{AB}} \approx 22.5$ mag, augmentation of the training sample has introduced the algorithms to the features associated with faint light curves. Comparing these results (Fig. 5.8) to those of previous training samples, we again see a significant improvement over the magnitude-limited training sample (Fig. 5.1). However, compared to the magnitude-limited plus faint training sample (Fig. 5.4), despite being much larger we do not reach the same classification performance. This is with the exception of ANN, which, as expected, does well when presented with large training samples (Goodfellow et al. 2016, Section 1.2.2), achieving higher AUC scores and more runs that reach 95% purity.

Going one step further, we augment the combined magnitude-limited and faint supernovae sample (from § 5.2.1), shown in Figs. 5.9 and 5.10 (MagLim+FaintAug). This differs from the previous augmented training sample as we now start with 'true' supernova light curves from fainter magnitudes, enabling the augmentation procedure to create more realistic faint light curves. The introduction of these produces our most successfully classified samples (ROC and purity curves in

[1]This is done using ASTROPY.COSMOLOGY.FLATLAMBDACDM with Hubble parameter $H_0 = 70$ kms^{-1}Mpc^{-1} and matter density parameter $\Omega_{\mathrm{m}} = 0.3$

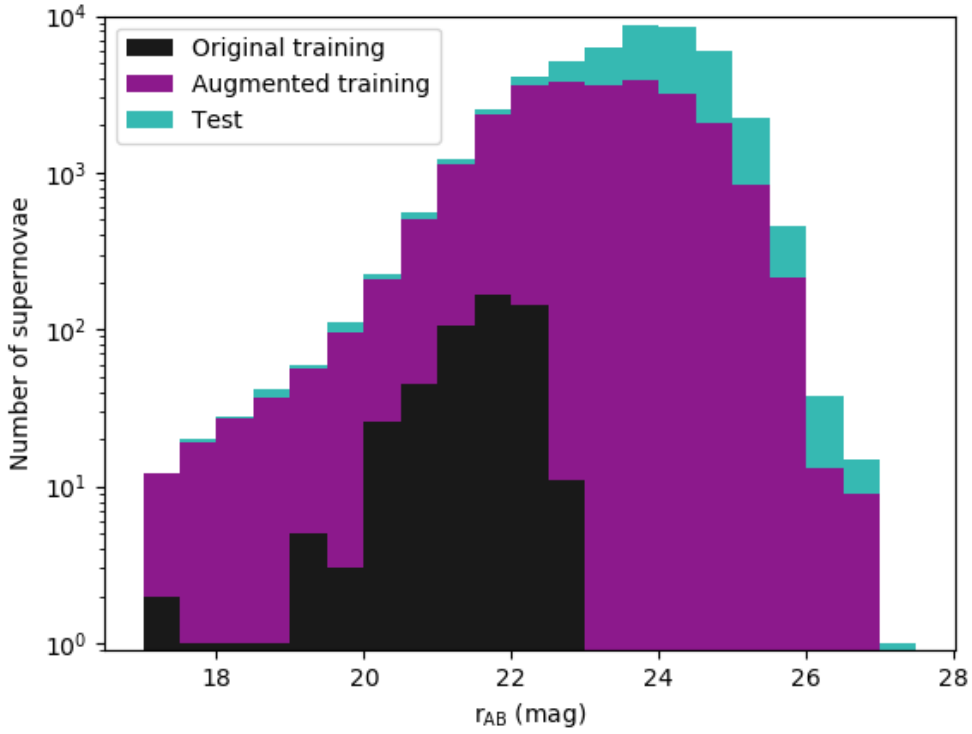


Figure 5.6: Magnitude histogram for the original training, augmented training (MagLimAug) and test samples. We augment the original magnitude-limited training sample from Fig. 4.6, increasing its size by a factor of 50 and extending it to much fainter magnitudes.

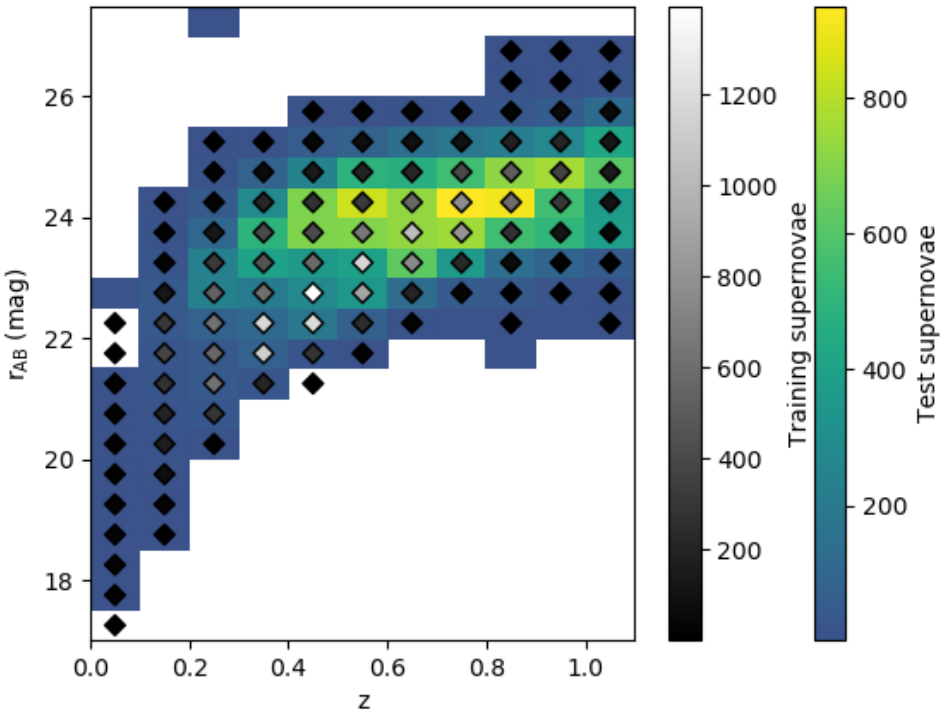


Figure 5.7: 2D histogram of the relative distributions of redshift and magnitude in the augmented training (MagLimAug; augmenting the sample from Fig. 4.7) and test set. Relative training and test numbers in each bin are not proportional, with more concentrated training supernovae at brighter magnitudes, although the training sample now covers the range of the test set.

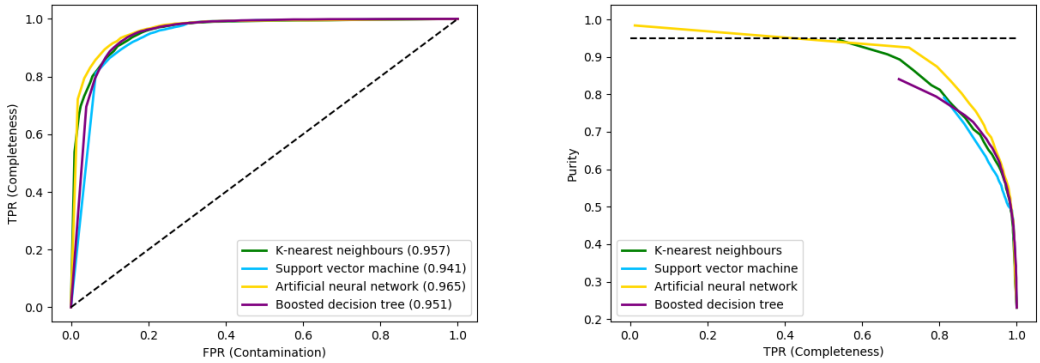


Figure 5.8: ROC and purity curves for the augmented training sample (MagLimAug). Augmenting the original magnitude-limited sample has a positive impact on the classification results.

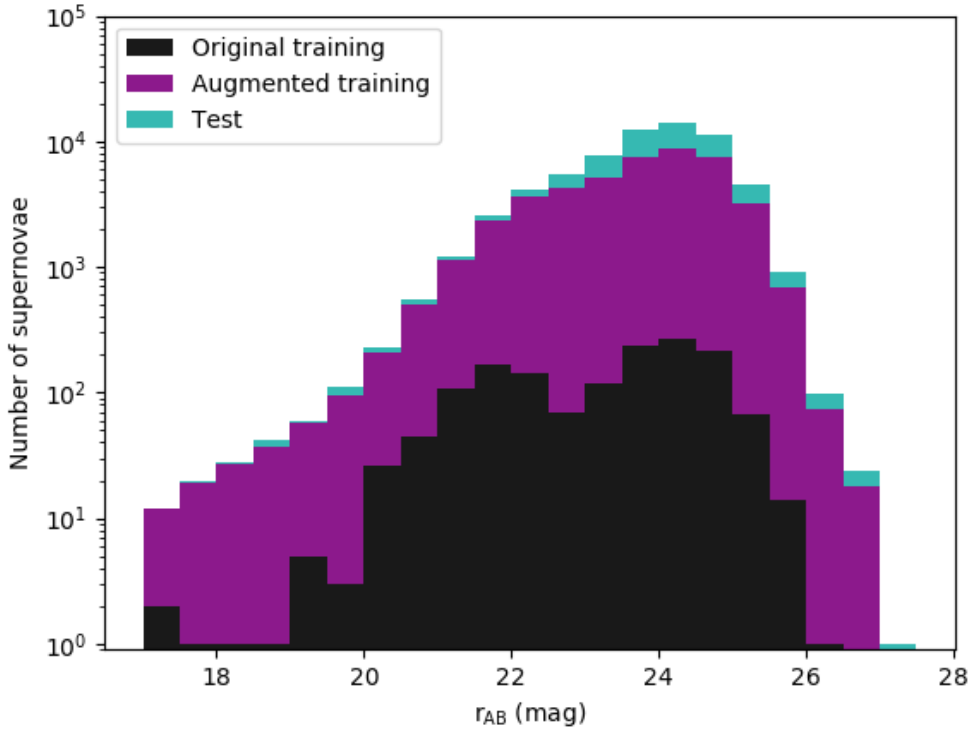


Figure 5.9: Stacked magnitude histogram of the original and full augmented training (MagLim+FaintAug; augmenting the sample from Fig. 5.2), and test set.

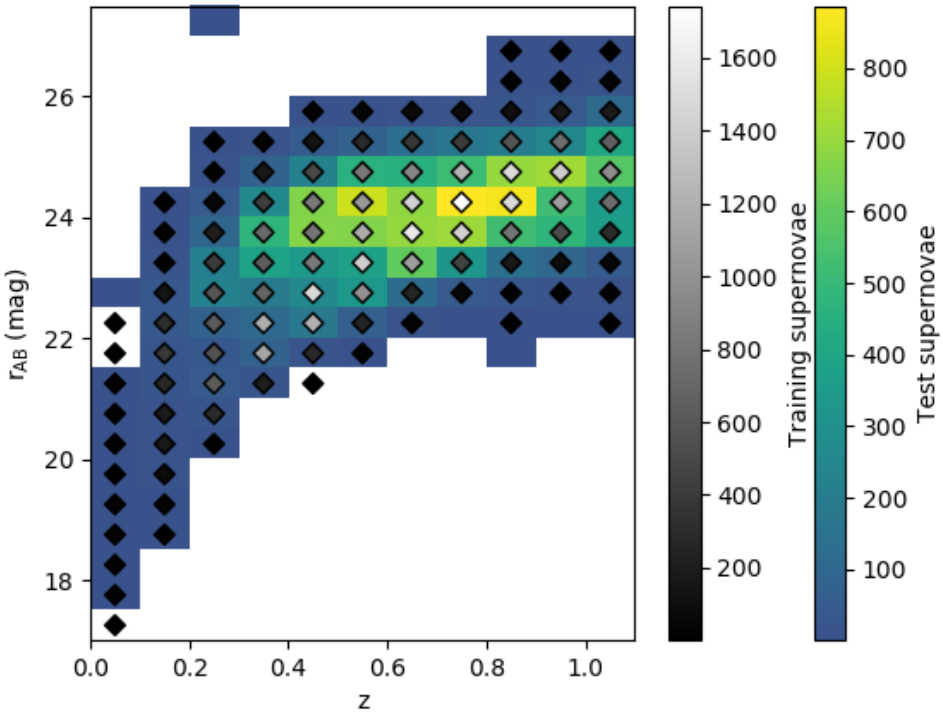


Figure 5.10: 2D histogram of redshifts and magnitudes for the full augmented original magnitude-limited training plus faint sample (MagLim+FaintAug). With full coverage of test set magnitudes and redshifts, there is also a higher concentration of fainter, high-redshift training supernovae than the previous augmented sample (Fig. 5.3).

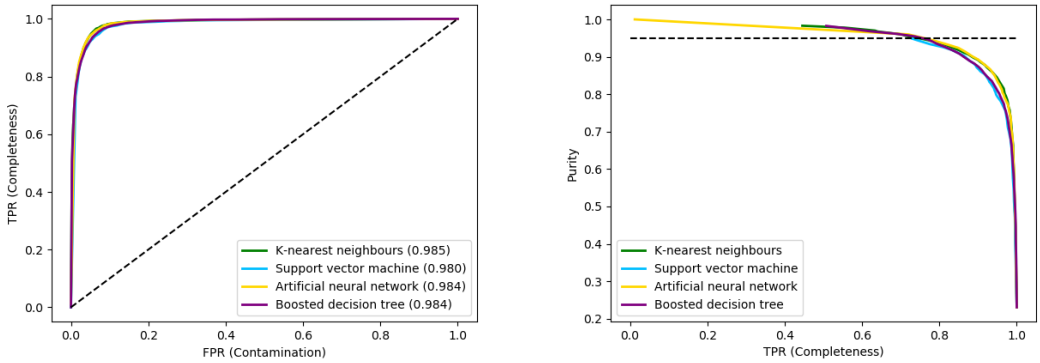


Figure 5.11: ROC and purity curves for the full augmented original magnitude-limited plus faint sample (MagLim+FaintAug). Augmenting the additional faint sample has improved further upon the augmented results from Fig. 5.8.

Fig. 5.11) and, compared to other training samples, is seen by the trend in the AUC boxplots in Fig. 5.12 and is summarised in Table 5.1.

In Fig. 5.13 we show the average AUC-dependence on redshift, comparing results for the original magnitude-limited training sample with the final augmented training sample (MagLim and MagLim+FaintAug). The large increase in size of training sample when augmenting has likely contributed to the effect of more predictable behaviour in the algorithms, shown by the very small error bars. Not only has augmented training improved the AUC scores at high-redshift, but also generally in low-redshift regions already covered by the magnitude-limited training sample. However, while consistently close to AUC = 1 at high redshift, further improvement is required for $z < 0.3$. Interestingly, the 0.0–0.1 and 0.1–0.2 bins for SVM actually performed worse for our most successful training sample.

5.3 Feature space

We show t-SNE plots comparing training samples (MagLim, MagLim+Faint, MagLimAug, MagLim+FaintAug) in Fig. 5.14 to help visualise why greater success is found with the addition of faint supernovae and augmentation. In each panel, the test set is the same, but the feature space covered by the training sample changes. The distribution and orientation of test set objects appear slightly

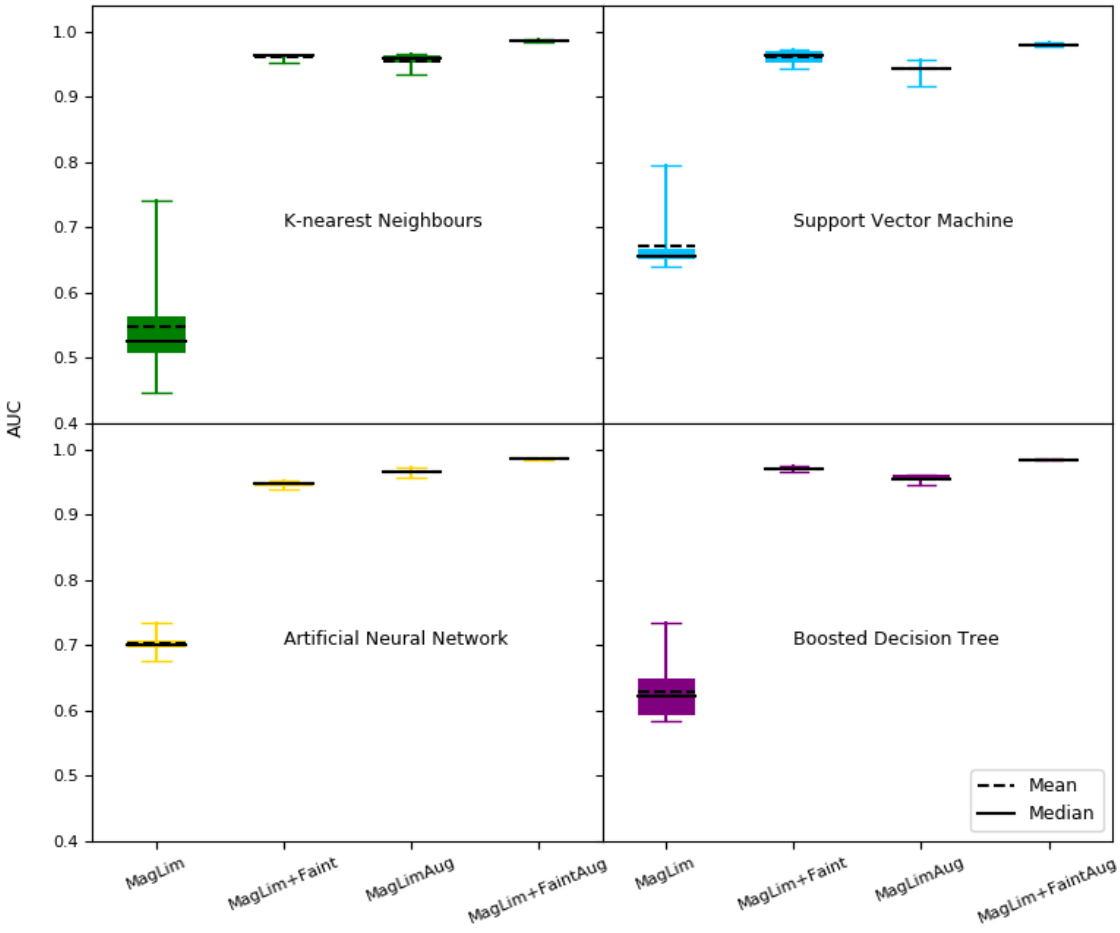


Figure 5.12: Boxplots showing the AUC scores over 10 runs for each of the four algorithms in four of our training sample simulations. The boxes represent the interquartile ranges, with their values shown in Table 5.1, along with means and medians. These results are for binary classification from Ia vs. non-Ia class probabilities. They are defined as MagLim: magnitude-limited training sample; MagLim+Faint: magnitude-limited sample with additional fainter supernovae; MagLimAug: magnitude-limited sample augmented; MagLim+FaintAug: combined magnitude-limited and faint samples both augmented.

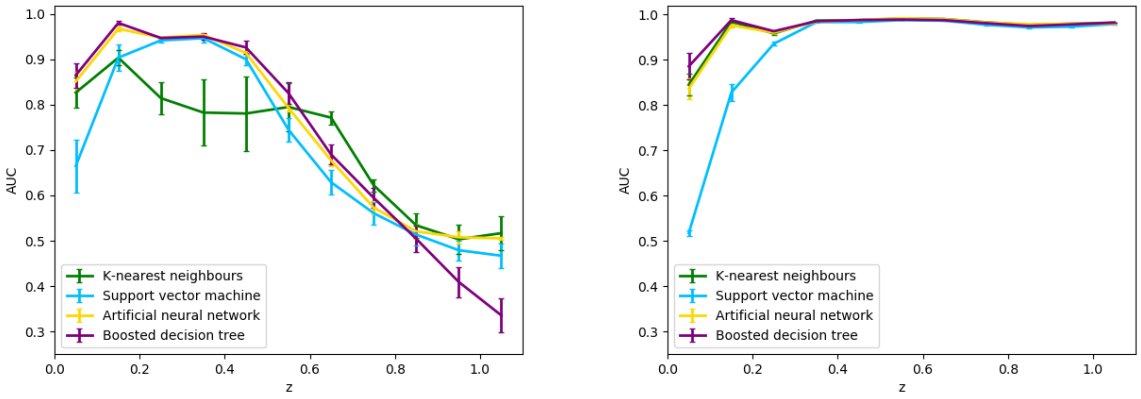


Figure 5.13: Average AUC scores as a function of redshift for all four algorithms, calculated in bins of size 0.1. Error bars represent the standard error in the average over the 10 runs. Left: MagLim. Right: MagLim+FaintAug

different due to the differing datasets considered (training and test together) and also t-SNE treating each plot with a random state instance. For magnitude-limited training, only a small region of the test set feature space is covered for both Type Ia and non-Ia supernovae, demonstrating why classification performance is not very successful. When we consider the proportions of different supernova types in magnitude-limited training (see next section), we find that the bias towards bright objects is also a bias towards Type Ia supernovae. Evidently, this is why there are so few training non-Ia supernovae in the feature space occupied by those in the test set. However, when the training sample is appended with faint supernovae and augmented, the training sample itself is not only much larger, but a significantly larger proportion of the feature space is now covered for the respective supernovae types, similar to the case for a representative sample as shown in Fig. 3.3. For MagLimAug, there are still some regions of feature space that are not covered, showing why this training sample does not perform as well as either MagLim+Faint or MagLim+FaintAug. This is likely due to AVOCADO's redshift constraints, preventing extrapolation far from where there is available data. It should be noted that the fractions of the supernova types in the augmented training sample are the same as those present in the original magnitude-limited sample, meaning that there remains a larger proportion of Type Ia supernovae

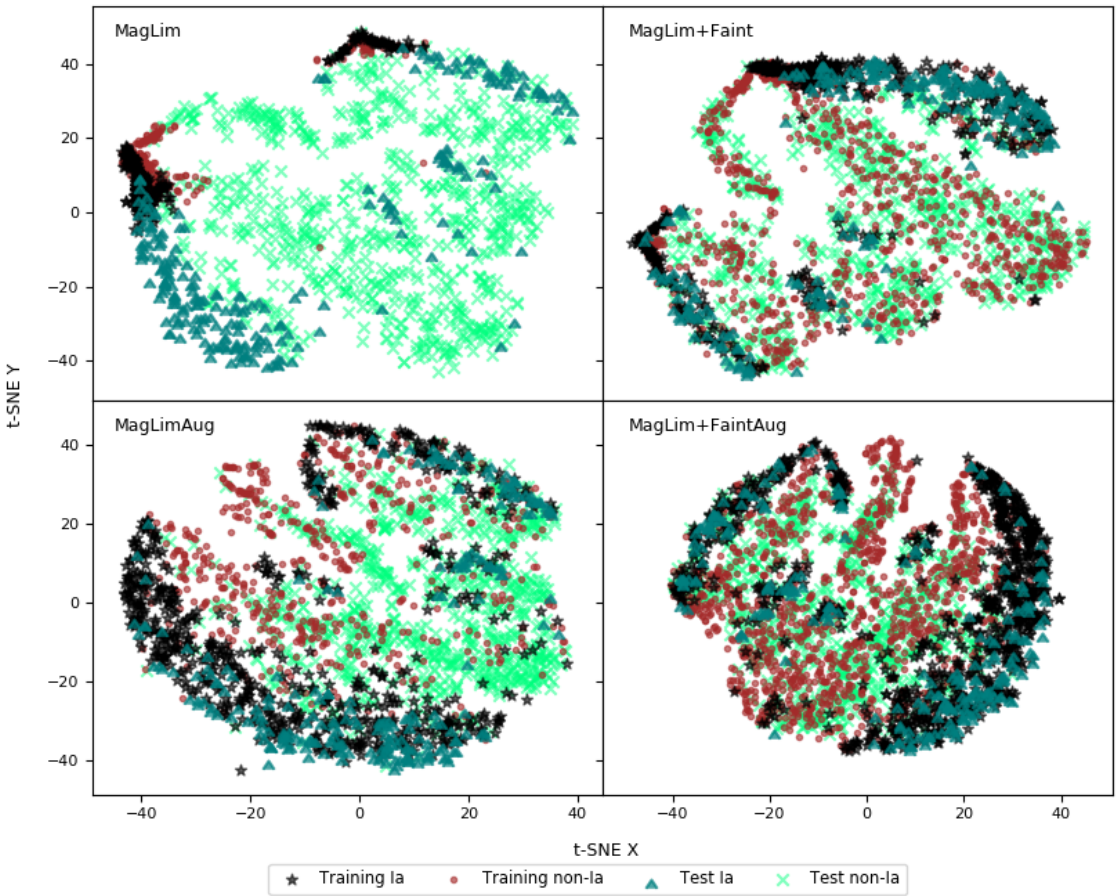


Figure 5.14: t-SNE plots comparing the feature-space coverage of four of our training samples. For clarity in these plots, one twentieth of the test set is shown for all, while one twentieth of the training is also shown for the augmented cases.

in the training sample than in the test set. This has relevance as the balance of classes in the training sample can potentially have effects on classification performance.

5.4 Class balance

When testing different training samples the balance of classes is not necessarily constant. If classes in the training sample are not balanced, then the algorithms may learn a bias towards the majority class. If the same imbalance exists in the test set, a class-preference may not be a negative side effect, but may help reinforce classification based on relative abundance of classes. However, it could also mean that the algorithm is simply better at identifying one class over another when, in general, we want a classifier to have the same degree of success when identifying any object type. Optimal class balance may appear arbitrary and may not reflect either balance in the training sample, or match with the natural distribution present in the test (Weiss & Provost, 2003). Some classifiers will forcibly alter the balance of classes, e.g. Möller & de Boissière (2020) creates a training sample with equal numbers of Ia and non-Ia supernovae.

Class balance is also important when it comes to assessing classification performance. A highly imbalanced training sample, e.g. 1 to 100, may achieve 99% accuracy by completely ignoring the minority class. Redefining metrics can point to where the problem lies and a change to sampling for the distribution of training classes will be required. This will usually be solved empirically (Chawla et al., 2004).

We ran several tests in order to analyse whether the balance of classes in training the SNMACHINE algorithms makes any noticeable difference to classification results. These are hypothetical tests to investigate the effects of balance, as we can never know the true balance of classes in a real test set. However, du Plessis & Sugiyama (2014) finds that by comparing the probability distributions of training and test data, an accurate estimate of the test set class balance can be determined.

Fig. 5.15 shows how the relative proportions of different supernovae types in the training and test sets change depending on how the training sample is

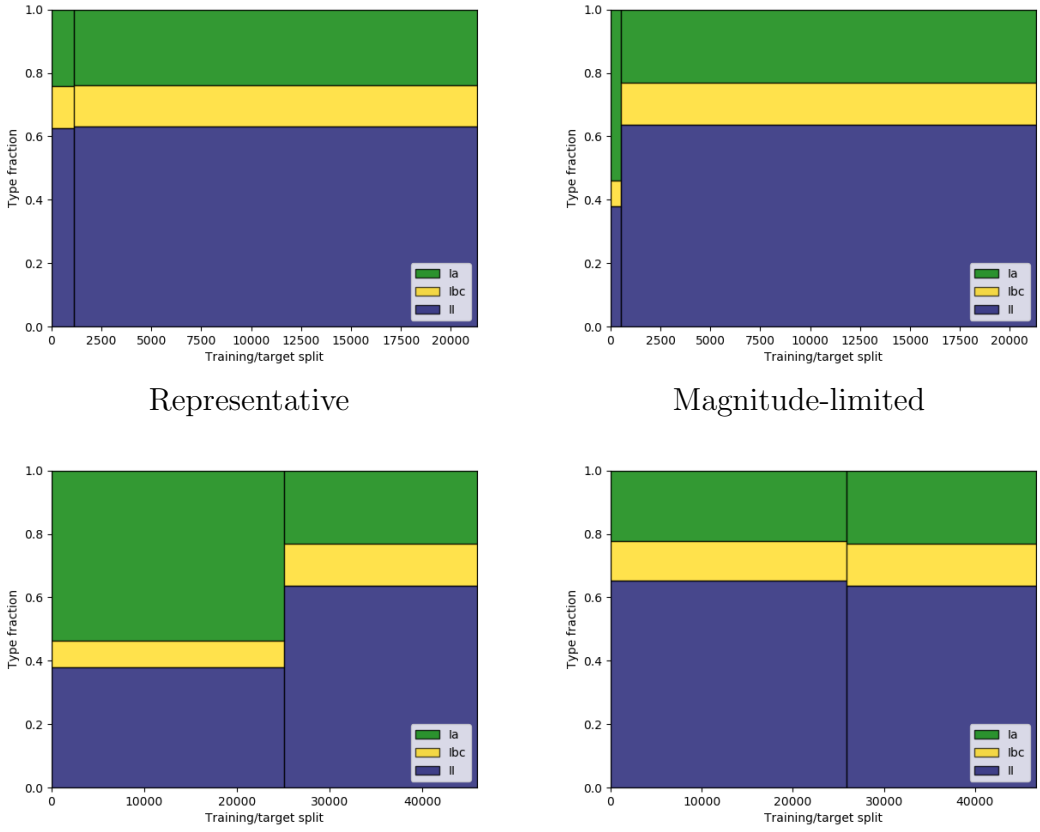


Figure 5.15: The split between different training and test sets, showing the proportions of different supernova types, grouped by Type Ia, Ibc and II. The vertical line in each plot separates the training (left) and test (right) sets. A representative training sample (size 1103) has proportions of these different types close to matching those in the test set. A magnitude-limited training sample (size ∼500) has a large bias towards Type Ia, as there is a higher proportion of Type Ia supernovae at brighter magnitudes. Augmented training has the same proportions of different types present in its original training sample, although is considerably larger in size. We can adjust the amount of augmentation per supernova type to match the balance of classes to the test set.

created. A representative training sample has similar proportions of classes as the test set. We ran a separate test to determine whether the poor results obtained when using a magnitude-limited training sample are simply because the training sample does not contain the same balance of classes as the test sample (because, for example, Type Ia supernovae are typically brighter than other classes). We fixed the proportion of Type Ia supernovae in the magnitude-limited training sample to match that of the test sample and find that there is no noticeable change in classification performance (AUC). The magnitude limit is causing some other features to be missing from the training sample, hence, to achieve accurate classification, success cannot be found by simply changing the balance of classes when the training sample is magnitude-limited.

To address the same bias towards Type Ia when magnitude-limited training is augmented, we ran a hypothetical test of augmenting training supernovae to match the balance of classes in the test set, whilst keeping the same total size (comparing bottom left and right in Fig. 5.15, which, along with the magnitude-limited case, have the exact same test set). Over 3 runs we saw a small increase in average AUC of 0.008, 0.011, 0.005 and 0.009 for KNN, SVM, ANN and BDT respectively. Applying the same technique to a training sample with faint supernovae included produced negligible change. In reality it will not be possible to know the exact proportions of different supernova classes in the test set. In practice, finding the optimal balance of classes in a spectroscopic sample would likely be a non-trivial task.

5.5 Beyond binary classification

We also ran tests in which the SNMACHINE algorithms are trained to recognise supernovae as being either Type Ia, Ibc or II, rather than the baseline Ia vs. non-Ia. This was done for the original magnitude-limited samples and the augmented magnitude-limited plus faint sample (comparing MagLim and MagLim+FaintAug to MagLim3Class and MagLim+FaintAug3Class respectively in Table 5.1). Considering mean AUC scores for MagLim, there is a small increase for ANN and

BDT although no significant difference is observed by making this change to classification.

For the augmented case, AUC scores are mostly very similar, the biggest change being a drop in average AUC of 0.05 for BDT. Also, quite notably for BDT, having 3 classes causes fewer runs to reach 95% purity - decreasing from 10 to 4. Conversely, SVM sees an increase from 3 to 10 runs whilst keeping AUC scores fairly consistent. In this 3-class scenario for our most successful type of training sample, it appears that SVM would be a better choice than BDT, although BDT would be more suited in all other cases that we tested. ANN also performs better with 3 classes, although this change is negligible; the AUC scores barely change at the third significant figure. No change at all is seen for KNN.

5.6 The optimised sample

Augmentation enables us to fill in some of the significant gaps in the test set feature space that we may not fully cover with our spectroscopic sample. Even though we have 'true' faint supernovae to help train the algorithms with improved representativity, when combining with the augmented magnitude-limited sample it is better to augment these as well, as shown in the results summary table (Table 5.1). Fig. 5.12 shows that we need these 'true' faint supernovae in our spectroscopic sample to achieve the highest AUC scores. There is a clear improvement in all algorithms going from purely magnitude-limited (MagLim) to adding faint supernovae (MagLim+Faint). As previously stated, the same positive trend is seen when these training samples are augmented (MagLimAug to MagLim+FaintAug).

We also compare these results to the hypothetical case of only having the faint sample and then augmenting that, with its results summarised in Table 5.1 as FaintAug. When we augment just the faint sample of supernovae, we get AUC scores very similar to those for MagLim+Faint (and MagLimAug). This seems to indicate that the original magnitude-limited sample may not be so crucial for training, as similar success is found by augmenting just a faint spectroscopic sample, however, we fundamentally do also need our magnitude-limited TiDES sample to obtain the best classification results (MagLim+FaintAug). Furthermore,

classification performance in the case of the augmented faint sample suffers due to the similar coverage issue of the original magnitude-limited sample, but at the other end of the brightness scale.

As an attempt to save on computing time and resources, we also consider the MagLimAug+Faint training, i.e. augmenting just the magnitude-limited sample and then adding the non-augmented faint sample of supernovae. However, compared to MagLim+FaintAug(3Class), this is not as favourable for classification in terms of AUC or purity.

With an original spectroscopic sample extending as faint as possible, these results highlight the important role of augmentation to achieve successful photometric classification in future supernova surveys. For this particular purpose, KNN, SVM, ANN and BDT all appear to be reliable machine-learning algorithms, reaching high AUC scores with very small variations, and also being able to achieve 95% purity over all test runs.

The work summarised in these last three chapters is published as a research paper in MNRAS (Carrick et al., 2021).

Table 5.1: AUC means, medians, interquartile ranges, maxima and minima for different types of training sample over 10 runs, and the number of those runs that reached 95% purity. The first four rows for each algorithm are the results shown in Fig. 5.12. We compare our results with additional training samples, including just the augmented faint sample (FaintAug). We investigate how SNMACHINE performs when returning 3 class probabilities (Ia, Ibc and II) for each supernova in the test set for magnitude-limited and augmented magnitude-limited-plus-faint samples (MagLim3Class and MagLimFaintAug3Class). Also, for the magnitude-limited case, we include results when using no redshift (MagLimNo-z). Finally, we include results from the runs investigating how adding the fainter supernovae (not augmented) on to the augmented magnitude-limited sample affected results (MagLimAug+Faint). We highlight in bold our most successful training sample for each algorithm, which is either MagLim+FaintAug, or MagLim+FaintAug3Class.

Algorithm	Training	Mean	Median	IQR	Max	Min	Purity 95%
KNN	MagLim	0.547	0.525	0.056	0.741	0.446	2
	MagLim+Faint	0.961	0.962	0.001	0.964	0.953	10
	MagLimAug	0.955	0.958	0.011	0.964	0.934	3
	MagLim+FaintAug	**0.985**	**0.985**	**0.001**	**0.987**	**0.983**	**10**
	FaintAug	0.973	0.974	0.004	0.980	0.958	10
	MagLimNo-z	0.631	0.640	0.053	0.707	0.536	0
	MagLim3Class	0.543	0.525	0.064	0.711	0.436	2
	MagLim+FaintAug3Class	**0.985**	**0.985**	**0.001**	**0.987**	**0.983**	**10**
	MagLimAug+Faint	0.976	0.976	0.001	0.978	0.972	10
SVM	MagLim	0.671	0.657	0.015	0.795	0.640	1
	MagLim+Faint	0.960	0.963	0.018	0.972	0.942	4
	MagLimAug	0.942	0.944	0.005	0.956	0.916	3
	MagLim+FaintAug	0.980	0.980	0.002	0.984	0.976	3
	FaintAug	0.953	0.960	0.009	0.965	0.914	6
	MagLimNo-z	0.617	0.601	0.088	0.714	0.504	2
	MagLim3Class	0.670	0.660	0.019	0.748	0.646	2
	MagLim+FaintAug3Class	**0.982**	**0.983**	**0.001**	**0.986**	**0.976**	**10**
	MagLimAug+Faint	0.968	0.969	0.004	0.972	0.958	1
ANN	MagLim	0.702	0.700	0.011	0.734	0.675	0
	MagLim+Faint	0.946	0.948	0.005	0.952	0.939	6
	MagLimAug	0.965	0.964	0.005	0.973	0.956	9
	MagLim+FaintAug	0.985	0.985	0.005	0.986	0.984	10
	FaintAug	0.957	0.962	0.027	0.978	0.926	9
	MagLimNo-z	0.613	0.614	0.060	0.688	0.568	0
	MagLim3Class	0.711	0.708	0.042	0.803	0.621	0
	MagLim+FaintAug3Class	**0.986**	**0.986**	**0.001**	**0.988**	**0.984**	**10**
	MagLimAug+Faint	0.978	0.977	0.002	0.982	0.976	9
BDT	MagLim	0.628	0.622	0.057	0.733	0.584	7
	MagLim+Faint	0.969	0.969	0.002	0.974	0.964	10
	MagLimAug	0.955	0.954	0.008	0.960	0.946	1
	MagLim+FaintAug	**0.984**	**0.984**	**0.001**	**0.985**	**0.983**	**10**
	FaintAug	0.956	0.959	0.007	0.974	0.922	10
	MagLimNo-z	0.642	0.635	0.062	0.709	0.560	6
	MagLim3Class	0.645	0.650	0.081	0.710	0.580	6
	MagLim+FaintAug3Class	0.979	0.980	0.002	0.981	0.977	4
	MagLimAug+Faint	0.976	0.976	0.001	0.978	0.974	4

Chapter 6

Contamination of the Spectroscopic Sample

Our work on optimising a training sample relied on the assumption that every object within the initial spectroscopic sample was correctly classified. Previous studies have investigated the implications of contamination in a photometric sample of supernovae for cosmological analyses (Jones et al., 2017; Vincenzi et al., 2021a), however, there has been little research done on the effects of a contaminated *training* sample when using machine learning for classification. This chapter discusses how accurate a realistic sample would be and the effects of contamination on the resulting photometric classification.

6.1 Choice of software

Since the start of the project there have been many new and innovative classifiers created (see § 3.1 for some examples), as well as much larger supernova datasets such as PLAsTiCC. In this part of the project we use a different classifier and a much newer dataset.

We use the software SuperNNova (Möller & de Boissière, 2020), due to its proven success and ability to deal with massive datasets using GPUs (Graphics Processing Unit). In recent years, GPUs have been much more preferable than

their CPU counterpart for running intense computations, particularly for running in parallel (Owens et al., 2008). Currently, the limiting factor is access to the GPU hardware, due to high price and high demand. We were able to gain access to the National Energy Research Scientific Computing Center (NERSC)[1] through DESC.

We use SuperNNova's deep recurrent neural network architecture for photometric classification of transients. Recurrent architectures are well suited for time-sensitive data, most commonly used for audio data (Mehri et al., 2017; Sutskever et al., 2014). This is because connections between hidden units allow a time-delay, controlling the flow of information to develop the model in such a way that it learns correlations across time. For this reason, supernova light curves are highly compatible with this type of network. As stated in Lochner et al. (2016) and Möller & de Boissière (2020), the choice of feature extraction method can significantly impact classification results. Therefore, the deep learning approach of SuperNNova is highly advantageous because feature extraction is not necessary, as it is in snmachine. The downside of this is that SuperNNova requires a lot of light curves to optimally train a model. Using simulations, there is no shortage of data and the issue becomes what is computationally possible.

6.1.1 Dataset

The data we use consists of simulated LSST transient light curves created using the `SNANA` package (Kessler et al., 2009). The dataset is similar to the DES simulations in Vincenzi et al. (2021b), but made to resemble LSST observations, and was created by others and provided by Maria Vincenzi (private communication). The transient observations' set of filters are u, g, r, i, z, y. The dataset objects are identified by their `SNTYPE`, an arbitrary number assigned to each class. Included types are 'normal' Type Ia supernovae (`SNTYPE 1`, SNIa-SALT2; Guy et al. 2007), peculiar Type Ias including SNIax (`SNTYPE 11`, SN2002cx-like; Foley et al. 2013; Li et al. 2003) and SNIa-91bg (`SNTYPE 12`, SN1991bg-like; Filippenko et al. 1992), SNIbc (`SNTYPE 20`), SNII (`SNTYPE 20`), SLSN (Superluminous Supernovae,

[1] https://www.nersc.gov/

`SNTYPE 70`), TDE (Tidal Disruption Events, `SNTYPE 80`) and CART (Calcium-Rich Transients, `SNTYPE 50`). Supernovae of Type Ibc and II were created using the core-collapse templates from Vincenzi et al. (2019). The total dataset includes a colossal ∼4.5 million objects, split by WFD and DDF observations, although we use only a small WFD subset of this to simulate a TiDES spectroscopic sample of approximately 30,000 objects. Exact values of the training, validation and test sample sizes used in our tests are 28,215, 17,634 and 17,635 respectively. This is a small training sample for SuperNNova and initial tests using default hyperparameters produced unfavourable classification (e.g. AUC $\approx$ 0.65 and accuracy $\approx$ 0.55). We therefore ran many tests in order to find an appropriate set of hyperparameters (discussed in the next section) for this dataset in order to achieve successful results on the order of success from Möller & de Boissière 2020.

We apply binary classification of normal Type Ia supernovae (SNIa-SALT2), vs. non-Ia, plus peculiar Ias including types Iax and Ia91-bg. From now on we implicitly include these when referring to non-Ia. One important aspect of the classification pipeline to note is that SuperNNova enforces class balance, so that the model is trained on an equal number of Ia and non-Ia objects. While it will not be far off from the true split of classes (TiDES anticipates ∼17,000 Type Ia supernovae from its ∼35,000 sample; Frohmaier et al., in preparation), this process artificially changes the balance from the dataset. Furthermore, we refer to plots tracing the training/validation stages and SuperNNova's metrics that assess classification on the final test set. These include AUC, purity and efficiency (TPR) as defined previously, and also accuracy, a metric that is suitable for when classes are balanced. Accuracy is defined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6.1}$$

and is a measure of the proportion of correctly classified objects, i.e correct identification of both the positive and negative classes, following the same Ia vs. non-Ia classification introduced in § 3.1.1. Since we are dealing with labelled data, the model is trained by supervised learning.

6.2 Hyperparameters in SuperNNova

The model is trained in a given number of epochs (not to be confused with light curve epochs), in which a training epoch consists of passing every training object through the network and, at the end of the epoch, the model is tested on the validation sample. Each individual epoch is split into batches, where we use the default `batch_size` of 128. The model's weights are updated after every batch. After all training epochs have been completed, the model is applied to a final test sample, a set of objects separate to the training and validation samples that the model has never seen before. The results of this final test produce the metrics (accuracy etc.) to evaluate the success of photometric classification.

We use the SuperNNova implementation from GitHub[1] and, based on initial testing, use the following hyperparameters values. These parameters are listed in the documentation.

By working on NERSC, we can utilise SuperNNova's GPU compatibility by adding the `use_cuda` argument. CUDA is a library that accelerates training of neural networks using GPUs (Chetlur et al., 2014). It is necessary as a way of configuring GPU parallelisation.

We set `data_fraction` = 0.2. This value has no significance on its own, but is used to produce a training sample of ∼30,000 objects from our subset of data. In reality we may not have as many objects with spectroscopic classification for validation, and our test sample is likely to be much larger as well (all LSST transients that have a spectroscopic host-galaxy redshift).

We include spectroscopic redshift information for the reasons discussed in § 4.4.3. This is done by setting `redshift` as `''zspe''`.

The `learning_rate` is set as 10^{-4}, one tenth of the default value. This was chosen as it produced a model that smoothly converged to a training/validation plateau without resetting. However, the learning rate is not such a simple parameter to choose, since it affects the amount that the model's weights are updated after each training batch and requires fine-tuning to achieve a balance between the model overfitting and not learning enough. Accuracy and loss during training are illustrated in Fig. 6.1 to show this. Indeed, in this case, the validation

[1] https://github.com/supernnova/SuperNNova

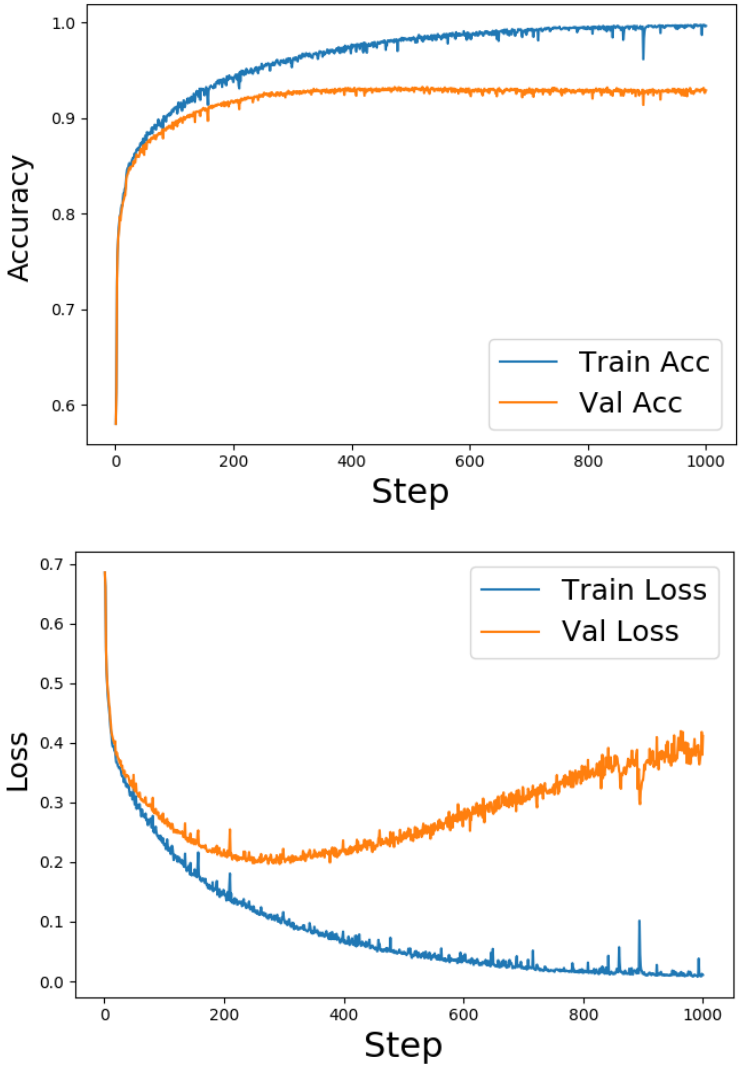


Figure 6.1: Accuracy (top) and loss (bottom) calculated after each training epoch (step) for the training and validation samples. With `learning_rate` set to 10^{-4}, the model's accuracy converges smoothly, shown by the curve flattening. However, discrepancy between the training and validation implies overfitting as the model does not generalise well to data other than the training sample. Significant overfitting is seen in the validation loss that increases before the model is fully developed: given these settings, training is optimised around epoch 250 (minimised loss), although at this point accuracy has not plateaued and loss is still relatively high.

loss increased significantly before the training loss had minimised over many epochs. Furthermore, to produce this model we used 1,000 epochs. This amount is much higher than should be necessary (Möller & de Boissière, 2020), and takes ∼3.5 hours. Hence, rather than simply choosing a learning rate, we use cyclic learning.

SuperNNova has the ability to apply cyclic learning. This innovative method varies the learning rate in cycles, reaching an optimised model much faster (Smith, 2017). Using cyclic learning, the model is trained in as few as 30 epochs, changing phase at epochs 10 and 20, and taking <10 minutes. This is presented in Fig. 6.2. At least the same levels of accuracy are achieved as our previous example in Fig. 6.1. Overfitting is still present but to a much lesser extent.

Finally, we apply the `cosmo` normalisation to the data. As we use redshift information, this normalisation is used to remove distance information of Type Ia supernovae by normalising each light curve to 1, so that any bias in classification for cosmology is removed.

6.3 Testing contamination

6.3.1 Changing object types

To test the effects of a contaminated training sample on final classification we run tests on the same selection of objects, first with a perfect spectroscopic sample, and then with a contaminated sample following the error rates of DASH (Deep Automated Supernova and Host classifier)[1] presented in Fig. 6 of Muthukrishna et al. (2019b): 1% of Ias (normal) are misclassified as Ia-91bg and 16% of Ia-91bgs are misclassified as Ia. For Ibc it is a bit more complicated as Ibcs are grouped in our dataset. Two out of the six DASH subtypes (Ib-norm and Ic-broad) have 6% chance of being misclassified as Ia, so, assuming Ibc subtypes are equally populated, 2% of all Ibc are misclassified as Ia. For misclassifications between Ibc and II, nothing was changed as these types both use the same identifier value

[1] This assumes that we use DASH to classify TiDES spectra. DASH itself uses deep learning by training on classified spectra from the literature, which may also contain errors. This effect should be looked at in the future.

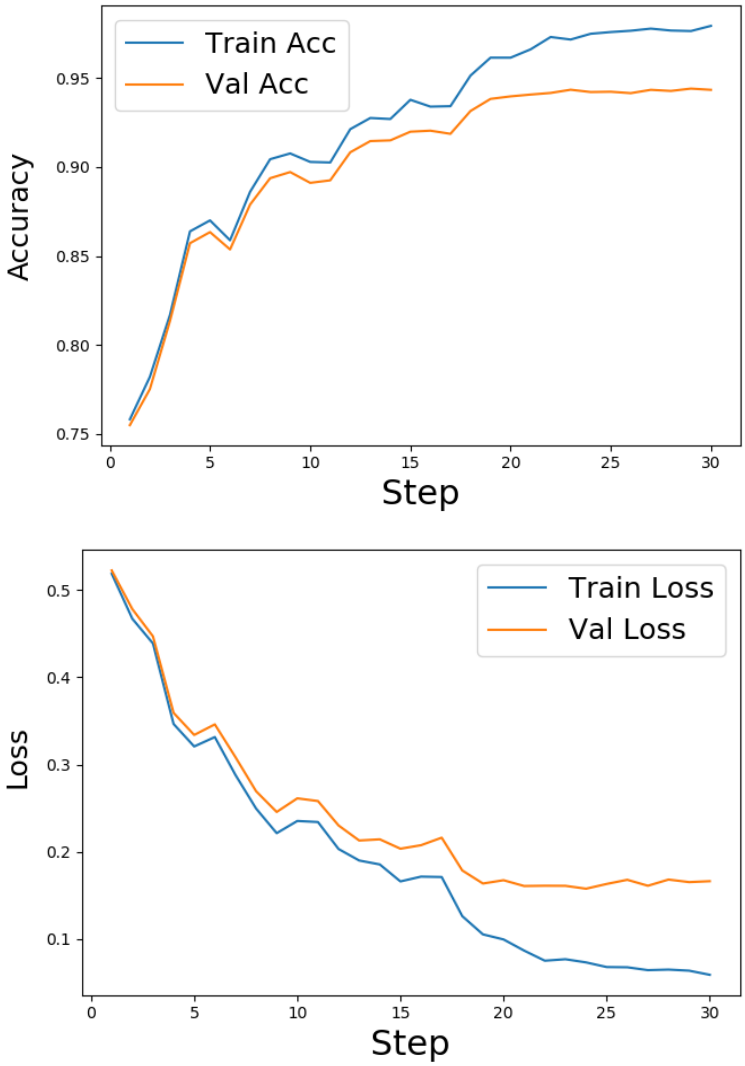


Figure 6.2: The same as Fig. 6.1 but for cyclic learning. Using cyclic learning, we can achieve more successful training with as few as 30 epochs.

(`SNTYPE 20` as discussed in § 6.1.1) in the simulated dataset. Furthermore, they both also exist together in the non-Ia class, so this type of misclassification does not affect the error rate of Type Ia supernovae.

We ensure that the test sample objects' classifications are not changed by checking against the list of `SNID` dataset objects that were not used in the original test set. Any other objects' classifications can be changed, and we do this by changing a random selection's `SNTYPE` with probabilities following the rate of misclassifications mentioned above (e.g. for a type Ia supernova, there is a 1% chance of changing it to a Ia-91bg and a 99% chance of it staying the same).

To use the altered copy of the dataset, an amendment was required in SuperNNova to make sure that the same objects were chosen for training/validation/test. Using the same seed value was not enough due to the dataset changes. Creating an output directory containing a copy of the file of `SNID`s used originally and removing SuperNNova's method to create this file produces the exact same training/validation/test split.

6.3.2 Results

As we use exactly the same selection of training objects in each test, any variation in our results comes from inherent randomness associated with neural network training. In particular SuperNNova's use of *variational inference* uses probability distributions to model weight uncertainty (Blundell et al., 2015; Gal & Ghahramani, 2016).

Following all training and validation stages, we show results of classification on a previously unseen set of test data in Fig. 6.3. Applying the contamination following DASH's error rates, we expected to see a decline in performance, but we do not see any significant decrease. Showing very similar levels of success, it appears that the network has chosen to mostly ignore the presence of false objects in the training, and instead focuses on the majority of what is included. We also tested a hypothetical scenario in which the training sample from our dataset has a 5% level of contamination, i.e. 5% (compared to 0.5% following DASH) of labelled Ias in the training sample are non-Ias and a similar number of Ias are labelled as either type Ia-91bg or Ibc. This test was carried out in order to see

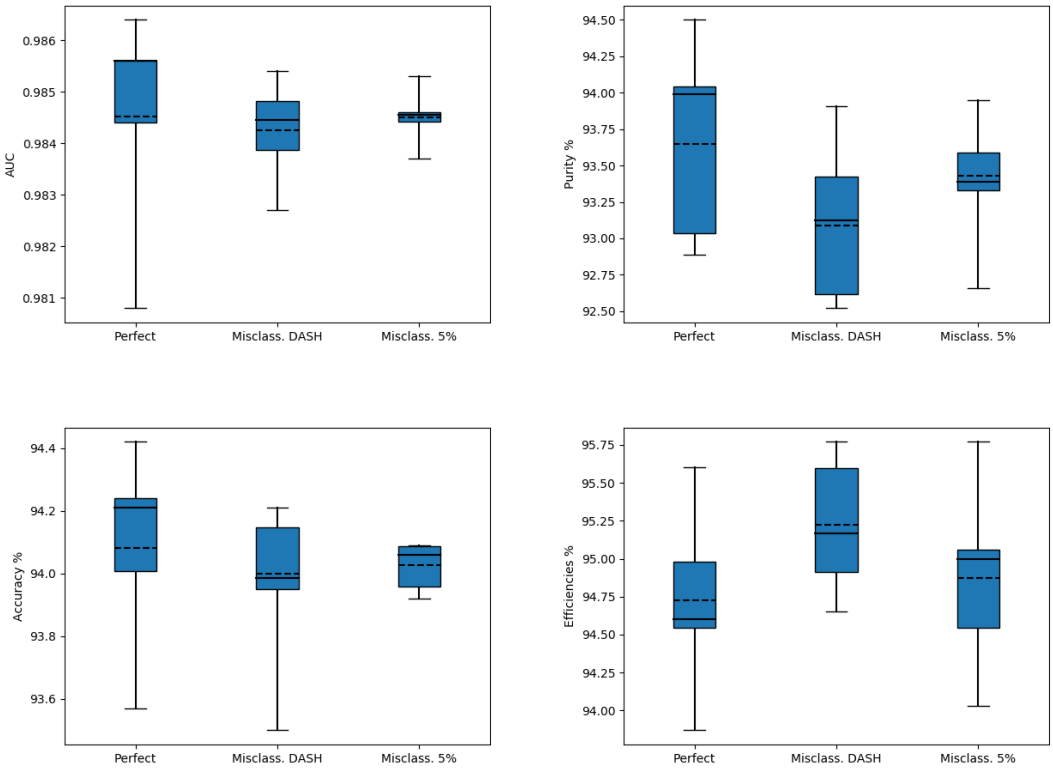


Figure 6.3: Boxplots summarising the results of 10 runs for the cases of a perfectly classified spectroscopic training sample, a contaminated training sample following the error rates from DASH and another sample in which the contamination is increased to 5%. Dashed line: Mean. Solid line: Median. Contamination appears to have little effect on classification performance.

whether a higher contamination would produce the expected drop in classification performance. However, even if we increase contamination to 5% we still see very little change from our test metrics: purity drops by less than 1% (from average 93.65%) and accuracy drops by 0.1% (from average 94.08%). AUC is the most constant score, only varying at the fourth significant figure: AUC = 0.9845 for both uncontaminated and 5% contaminated training, and AUC = 0.9843 using DASH levels of contamination.

Efficiency seems to be the odd one out of these metrics, as it experiences a slight increase (<1%, from average 94.73%) when using contaminated training. Furthermore, the spread of results reduces in all metrics apart from efficiency. These results indicate that the network is more likely to classify more objects of both Type Ia and non-Ia as Type Ia when the training is contaminated; it returns more correct Type Ia objects (increasing efficiency), while also incorrectly classifying non-Ia objects (reducing purity). This makes sense, as training the model with a contaminated sample makes it less sure about the true class of objects. Conversely, it is perhaps surprising that it seems to choose Type Ia as the preferential class as it contains less variation (only normal Type Ia supernovae, plus contamination), compared to non-Ia objects that have a much wider distribution of features.

The study Vincenzi et al. (2021a) found that the level of non-Ia contamination in a photometric sample ranges between 0.8–3.5%, depending on simulation models used. They demonstrate that cosmological bias due to a contaminated photometric sample is therefore small (bias on w of <0.008, and <0.009 and <0.108 for w_0 and w_a respectively) and can be mitigated further combining supernova data with CMB measurements. We find that a contaminated training sample makes little difference to the classification performance. An analysis through to the final cosmology has not been done yet, although, given these results, the outlook is very promising for upcoming supernova surveys.

Chapter 7

Unsupervised Classification in the ESO Archive

This chapter is dedicated to a separate data science project undertaken as part of my PhD studies. ESO is involved with ESCAPE[1] (the European Science Cluster of Astronomy and Particle physics) to improve data workflow and data-driven research. Using data from the ESO archive, we apply unsupervised deep learning in a separate classification study. With numerous type of astronomical sources from many different instruments, it is a big challenge to search for specific objects and ambiguity can arise from mistakes in object metadata, inconsistencies between different sky surveys and even a lack of thorough physical understanding of such objects.

To address this, ESO is developing deep learning approaches to classify data in its vast archive. The work predominantly takes place on *Rostam*, a computing cluster including four GPUs. Using the mantra 'letting the data speak for themselves', we use an autoencoder network to determine important features that characterise spectra. By compressing spectra and decompressing in a reconstruction task, the network is forced to create a low-dimensional representation of the data. Rather than relying on human efforts to judge how data in the archive should be classified, a deep learning approach allows the machine itself to determine what meaningful

[1] https://projectescape.eu/

features are hidden in the data. This differs from the machine-learning pipeline SNMACHINE introduced in § 3.1 as distinct wavelet features were extracted from the data to be used as input for the classification algorithms. In this project, the input is the data itself. Archive users will eventually be able to make more meaningful searches based on these features, i.e. what the machine defines as 'similar'. The 'classification' the network determines is learned without supervision from the low-dimensional representation of the data. This interpretation has some ambiguity, but also contains learned concepts of real astrophysics.

7.1 Convolutional autoencoder

An autoencoder architecture consists of an encoder, in which data is compressed into a low-dimensional representation (or 'code') and a decoder, which decompresses the data back to the original dimensionality. The decoder typically has the opposite structure to the encoder and both constituents can be constructed as one of many different types of neural network model, such as convolutional or recurrent. In this work we use a deep convolutional autoencoder. In convolutional layers the neuron weights are shared, depending on the size of the kernels. This process is significantly faster than treating each pixel individually and is typically associated with image recognition tasks (Simonyan & Zisserman, 2015). Convolution can be applied to spectra as, in general, adjacent pixels will be closely related due to the spectral continuum and spectral features may be found across multiple pixels.

We present a detailed diagram of the network in Fig. 7.1. The input data, 1D vectors of flux measurements (spectra from the ESO archive), are transformed down to their low-dimensional representation through 15 convolutional layers and one final fully-connected layer. Through another fully-connected layer and 15 up-convolutional layers, the code is transformed back to the original spectrum dimensionality. The network determines what information in the data is important by being trained to reconstruct the original input spectra. This reconstruction produces the output comparison to the input spectra, which is required to determine a loss for training the model.

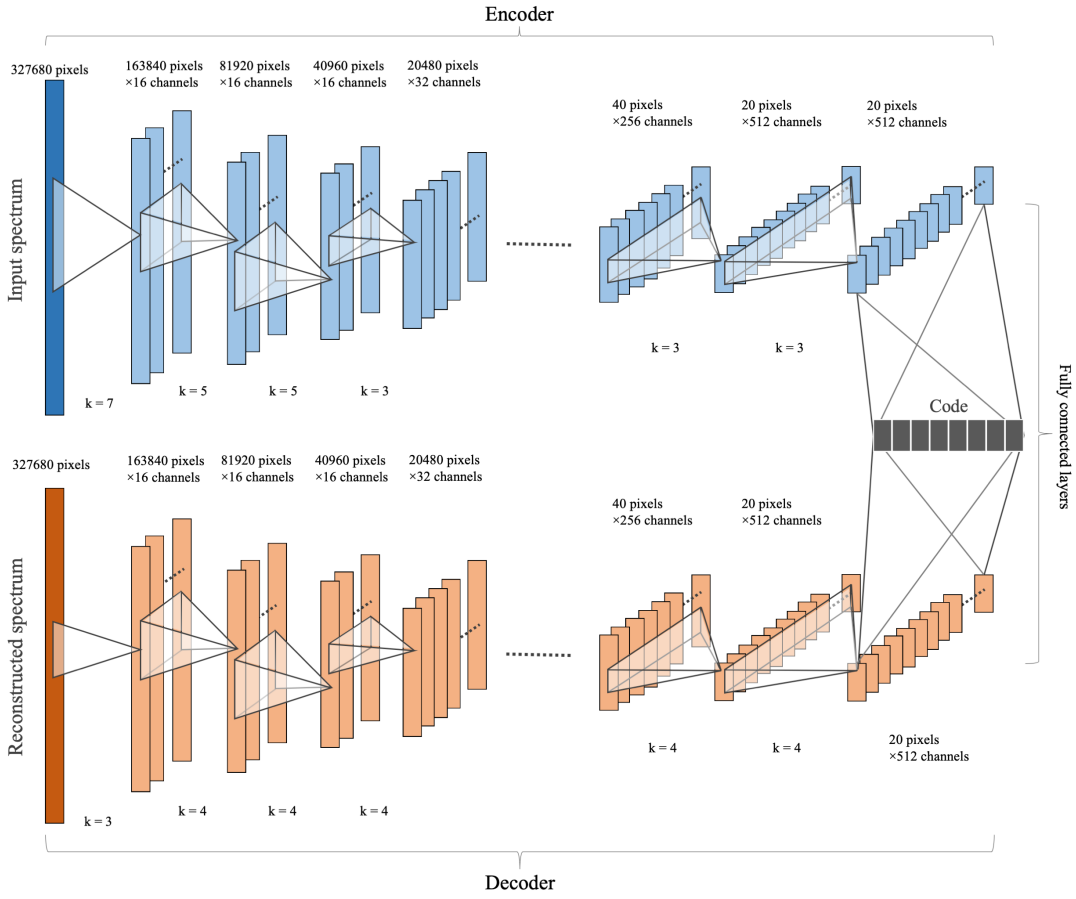


Figure 7.1: Detailed architecture of the deterministic autoencoder. Due to lack of space, not all the layers have been visualised. The 'k' values given are the kernel sizes.

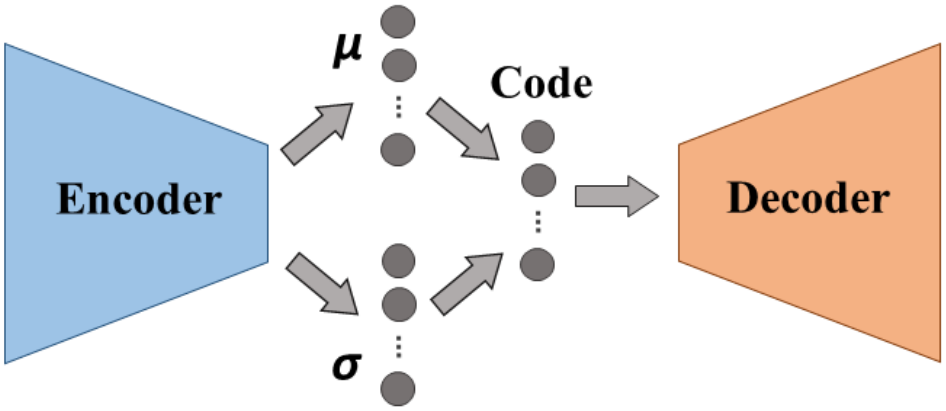


Figure 7.2: The variational autoencoder architecture. This is identical to the deterministic autoencoder in Fig. 7.1, with the exception that the code is not directly connected to the encoder, but is drawn from learnable parameters of a normal distribution.

Two autoencoder types were tested: a deterministic autoencoder, following the description above, and a non-deterministic variational autoencoder, in which the code is not directly connected to the encoder, but consists of values drawn from a normal distribution defined by a mean and standard deviation for each latent variable learned by the network (Kingma & Welling, 2014). A simple schematic of the variational counterpart is shown in Fig. 7.2.

After combining the weights at each network connection, the strength of the neuron output is determined by a learnable parameter constrained by an activation function. Neurons that provide the required flow of information for reconstruction are set as active through training. In this network, neuron outputs are calculated using a Leaky ReLu activation function, which has a non-zero gradient over its entire domain (Leaky Rectified Linear Unit, Maas et al. 2013); it allows a small leak in signal even when the neuron is not active.

The model automatically learns features of the data by passing it through the information bottleneck (Tishby et al., 1999). Each node in the latent space is assigned a feature that contains information necessary for reconstruction. This is an example of *representation learning* (Bengio et al., 2013). Training the network model to reconstruct spectra is done by optimising a pixel-level accuracy through

minimisation of the L1 loss function, which is empirically calculated as

$$\mathcal{L} = \frac{\sum_{i \in \mathcal{M}} |x_i - \hat{x}_i|}{n}, \tag{7.1}$$

where i is the pixel index, x and $\hat{x}$ are the original and reconstructed spectral flux respectively, and n is the total number of pixels. $\mathcal{M}$ is a mask which only considers pixels that we are interested in, i.e. it ignores the 'zero-regions' defined during the data preparation (see § 7.2). Loss is monitored via testing with a validation sample every epoch and the training progress can be saved at any epoch to continue developing the model later. Training and validation subsamples are split after sorting by each spectrum's `ADP ID`, a unique identifier that has no physical bearing on the spectrum. Once the loss has stabilised, the network model has converged and cannot be improved upon further without changing its training sample or architecture, which would generally require developing a new model from scratch.

7.2 Preparing the data

The archive data we used were observations consisting of 272,376 stellar spectra obtained using the HARPS (High-Accuracy Radial velocity Planet Searcher) instrument, a high-resolution spectrograph dedicated to the discovery of exoplanets (Pepe et al., 2002).

Before training, it is essential to have data that is homogenous, so that the network does not learn to interpret inconsistencies such as instrumental effects. Three 'zero-regions' are included at the start, middle and end of each HARPS spectrum. These zero-regions were defined based on the wavelength limits of non-zero flux regions when considering all spectra in the dataset. This ensures that all spectra have the exact same start and end wavelengths, and also have the same gap in the middle, an instrumental characteristic present to some extent in all dataset objects. Therefore, a constant mask covers all the spectra at the cost of losing a small fraction of pixels, and instrumental artefacts are removed whilst the astrophysical interpretation of the spectra remains.

It is also necessary that the spectra have exactly the same dimensionality. The total spectrum size was chosen to be 327,680 ($2^{18} + 2^{16}$) for computational purposes, as each time the data is compressed via network layers its size is reduced by a factor of 2, and also because this number of pixels covers the whole spectral range whilst adding little unnecessary padding either side of the non-zero regions. Each pixel of the spectra represents a specific wavelength. Using pixel indices to represent wavelengths means the input is only a 1D vector of flux values, a much simpler dataset to deal with than combined wavelength and flux data. The wavelength range of the spectra is 3,785Å to 6,910Å.

Additionally, we remove 'unstable' spectra, defined as those that contain undefined (`NaN`) or unrealistic (e.g. $> 10^8$ adu) flux values, reflecting instrumental errors.

Preparing the data is arguably the most essential step prior to training a network model, as lack of homogeneity could result in chaotic interpretations with learned features that are not helpful in understanding any of the underlying astrophysical phenomena.

7.2.1 Unique objects

In order to investigate the network's interpretation of different objects, we want to study the learned latent representation with respect to a sample of 'unique' HARPS objects, i.e. a sample in which each HARPS object appears only once. This presented its own challenge, as the HARPS dataset includes many different surveys that do not necessarily use the same name identifiers for the same objects. Furthermore, there are several instances of objects being assigned incorrect meta-data, such as name typos. There are also observations of solar system objects such as the Sun, Moon, Jupiter and its Galilean moons, and asteroids. As these are not stellar spectra characteristic of HARPS' main objective, they contaminate the sample. However, they were left in to keep the degree of supervision close to zero.

Several approaches were tested to make an algorithm that produced a set of true unique objects. We tried several different clustering algorithms such as nearest neighbour searches including `sklearn.neighbors`' `KDTree` and `BallTree`

algorithms, and also `sklearn.cluster.DBSCAN`, a density-based clustering. The aim was to identify unique objects based on their cluster of sky coordinates, as observations of the same object should appear at the same points in the sky. However, a flaw of this method is due to solar system objects such as those previously mentioned, especially the Sun and Moon, whose positions in the sky constantly change. These objects would always be assigned to multiple different clusters. The main difficulty lies in fine-tuning the clustering method, especially as the number of observations of different objects is very inconsistent: some objects may have one or two observations while, e.g. αCen-B has ∼20,000.

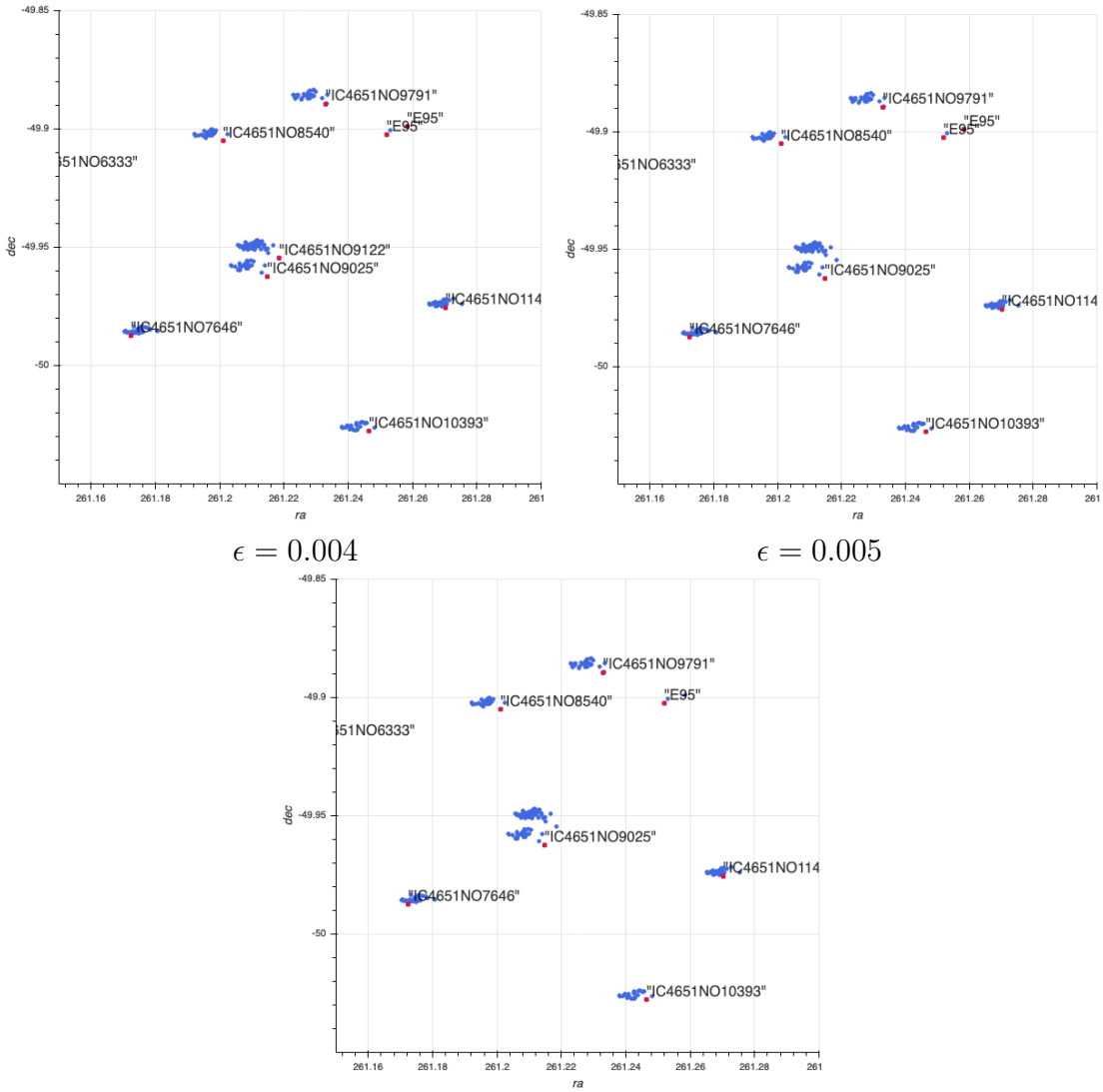
DBSCAN's ϵ parameter is defined as a limit on distance for two samples to be considered in their mutual neighbourhood, but is not necessarily a limit on distance between points in a single cluster. The 'elbow method' was used to determine the optimal ϵ parameter. Using Bokeh[1], an interactive tool to visualise results of clustering, we show an example of why this method is insufficient in Fig. 7.3 to determine unique objects. By cross-checking with additional metadata, we notice that that increasing ϵ from a low value may, in some areas, combine observation clusters of multiple objects into one before it correctly identifies single low-density clusters. Hence, a perfect solution could not be found and it became a case of balancing between over- and under-clustering. Initial tests made it apparent that it was impossible to obtain a set of unique objects without considering their object names at all.

In the end, we defined our set of unique objects based on their 'fixed' names: the objects' names were made upper-case and any dashes, underscores or points were removed. This results in a sample of 7,653 unique objects. We accept the contamination and potential name errors in this analysis as it can only affect the results negatively, producing a lower-bound to success rather than being overly-optimistic.

7.2.2 Latent space

The latent space is the lowest-dimensional representation of spectra in the data, sandwiched by fully-connected layers joining the encoder to the decoder parts

[1] https://bokeh.org/



$\epsilon = 0.004$ $\epsilon = 0.005$

$\epsilon = 0.006$

Figure 7.3: A snapshot in R.A. and dec. of some of the HARPS data for different values of ϵ. Blue points are individual observations and red points represent a single observation within a cluster determined by the DBSCAN algorithm. For $\epsilon = 0.004$ the object "E95" is split into two clusters. Increasing to $\epsilon = 0.005$, this is still the case and the two observation clusters of objects "IC4651NO9122" and "IC4651NO9025" have incorrectly been identified as one cluster. At $\epsilon = 0.006$ "E95" has been correctly identified as one cluster, although at the expense of incorrect clustering elsewhere.

of the network. Reconstruction quality is highly dependent on the number of dimensions in the latent space. In the VAE, disentanglement is used to ensure that each node of the latent space has as close as possible to its own unique significance in representation of the data. A fine balance of disentanglement is required as: too little and the interpretation of the data is spread across too many dimensions to understand individually, and multiple latent nodes may be assigned the same features; too much and the reconstruction suffers as the finer details are completely overlooked and the network cannot preserve details any longer as it is too simplistic to reconstruct spectra and learn informative information. We want each latent node to be 'orthogonal', where the features are independent of each other (Burgess et al., 2018). In the variational autoencoder, the amount of *disentanglement* can be controlled by λ, a training parameter and weighting that is used to enforce orthogonality between features (Tschannen et al., 2018). We find that $\lambda = 0.3$ is an optimal value for this task, where no two significant dimensions show significant correlation, implying a sufficient level of disentanglement.

The number of latent dimensions also has a significant impact on the reconstruction quality of spectra. Dimensions between 2 and 128 in powers of 2 are tested. We present results for the baseline 2, and also 8 and 128 which exhibit similar behaviours in their latent information (the other networks do not contribute other findings unique to our analysis). Examples of reconstruction of a spectrum are shown in Fig. 7.4 using these latent spaces for both the autoencoder and variational autoencoder. As mentioned above, disentanglement comes at the cost of reconstruction quality, and a higher number of latent dimensions is required to compensate and produce adequate reconstruction.

7.3 Physics learned by the network

To determine what the network has learned, we compare its low-dimensional representation of the data to physical tags for every possible dataset object, cross-matching `ADP IDs`. We utilise a set of available astrophysical and observation-time features including effective temperature, surface gravity, radial velocity, airmass

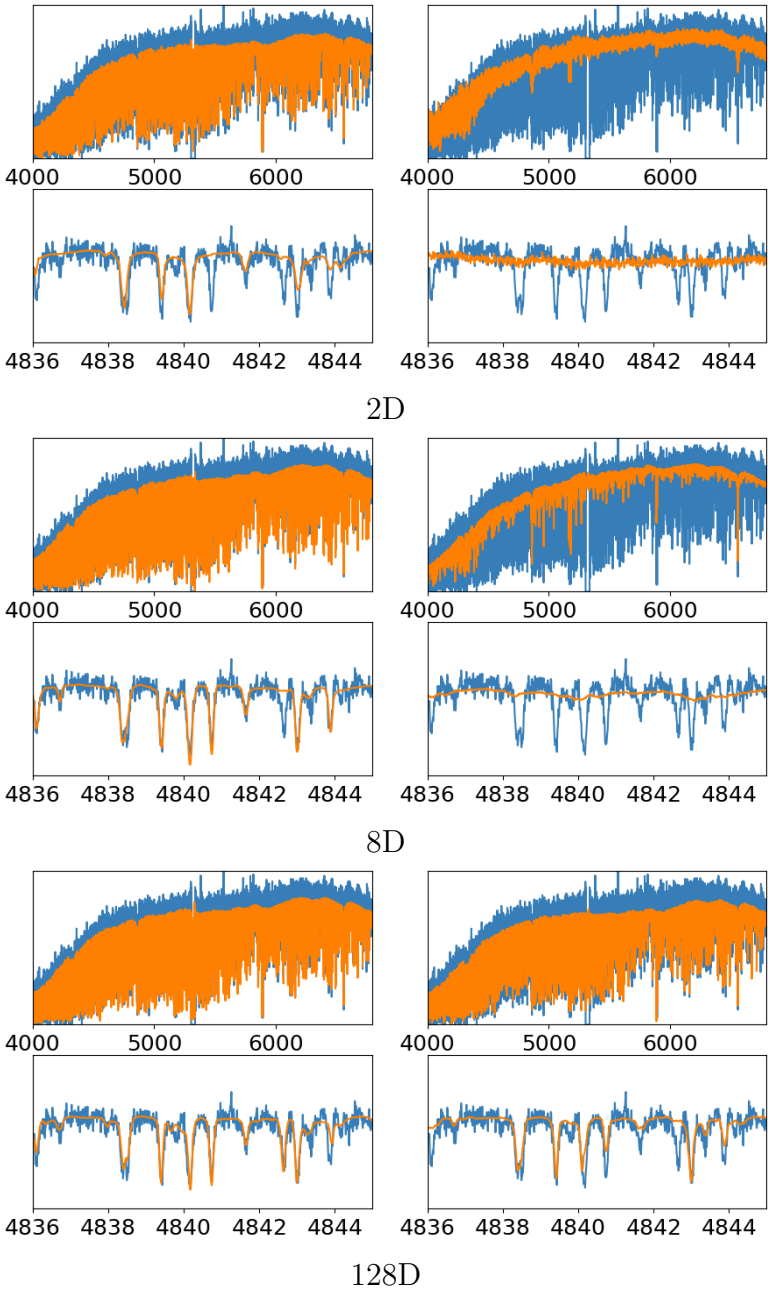


Figure 7.4: Reconstruction of an example spectrum for both the deterministic (left) and disentangled variational (right) autoencoders for different latent space dimensions, labelled 2D, 8D and 128D. The blue spectrum is the input and the orange spectrum is the output reconstruction. Reconstruction quality is reduced for the sake of disentanglement, and improved when latent dimensions is increased.

and signal-to-noise ratio from SIMBAD (Wenger, M. et al., 2000) and the TESS Input Catalogue (Stassun et al., 2019).

Interestingly, with $\lambda = 0.3$, latent dimensionalities of both 8 and 128 yield six informative dimensions. The measure of informativeness of nodes is given by the median absolute deviation (MAD) of statistical dispersion. The logic behind this is that dispersion represents underlying variability across multiple objects. When using a latent space of 128 dimensions, we find that the autoencoder has assigned specific nodes to radial velocity and effective temperature. The node assigned to effective temperature also correlates with surface gravity, due to the close relation between these two physical parameters; a linear correlation between $\log(g)$ and $\log(T_{\rm eff}$ is found following both parameters being determined through photometry (e.g. *uvby* β, Napiwotzki et al. 1993). The remaining 4 informative dimensions do not correlate with physical parameters, but leave room for future study into what exactly makes them 'informative'.

Fig. 7.5 shows the relationships between node 85 of the latent space, and the effective temperature $T_{\rm eff}$ and surface gravity $\log(g)$. Colour-coding the datapoints by spectral type of the stars helps to illustrate the astrophysical variability with respect to these parameters and how they are picked up by the network. The correlation between this node and effective temperature appears much tighter, although the fact that there is correlation with two astrophysical parameters is due to the shared information between effective temperature and surface gravity of stars, and the network does not need to dedicate separate nodes for these different physical characteristics.

Fig. 7.6 shows the relationship between node 124 and radial velocity. The network has clearly learned the concept of zero-velocity and formed a symmetric function around this point. However, including spectral type information reveals that the correlation is stronger for cold stars, while for hot stars the correlation vanishes. This is possibly due to increasing sparseness of absorption features with increasing temperature.

Through reconstruction training, the network has determined what it considers important information encoded within the stellar spectra. This is effectively the autoencoder's own classification of the data, and includes notions of real astrophysics along two of its informative latent dimensions. The network's perception

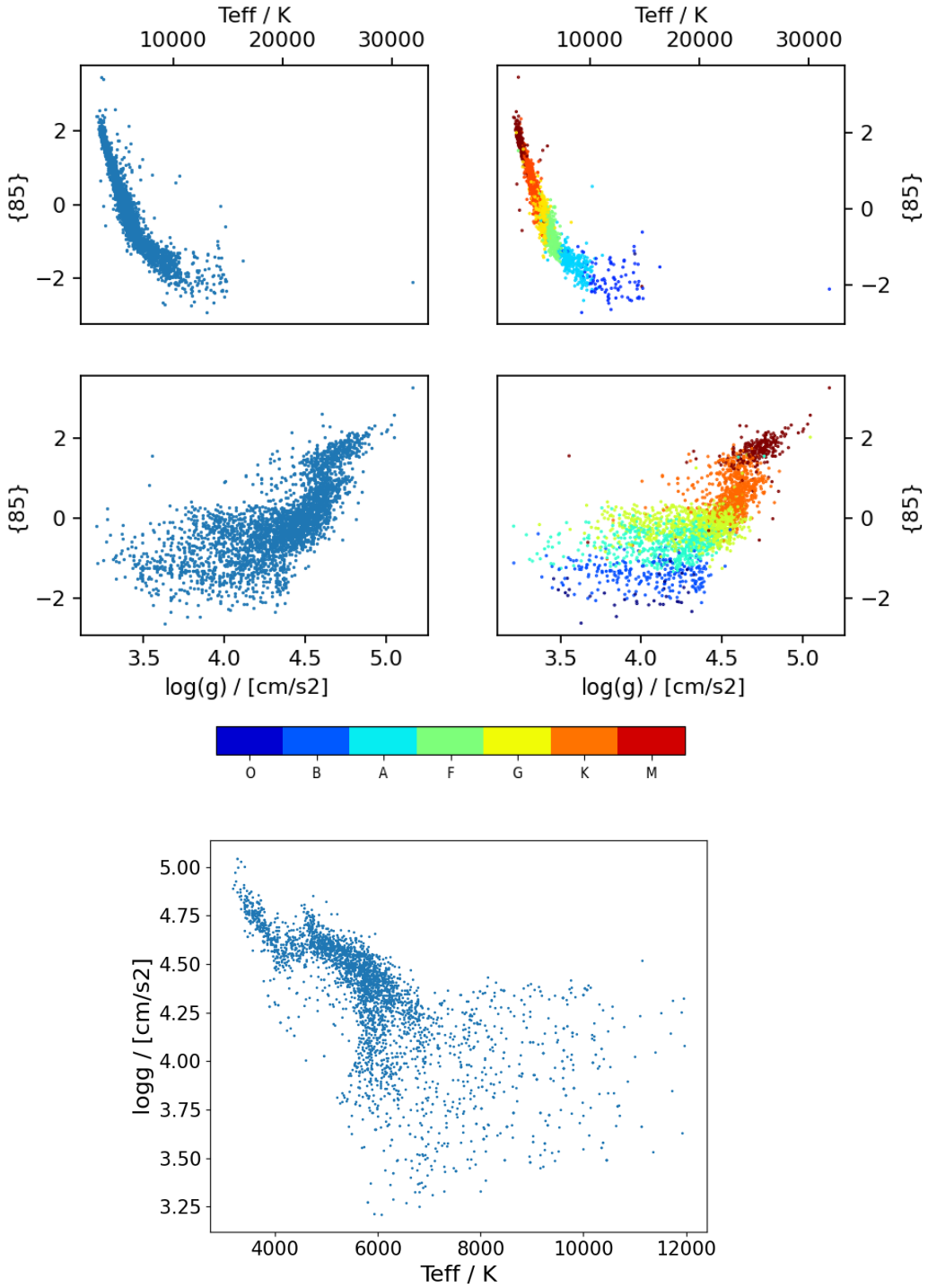


Figure 7.5: Correlations between node 85 values and T_{eff} (top row), and $\log(g)$ (middle row). In the right panels, the data points are colour-coded by the spectral type of the star, as given by SIMBAD. On the bottom we show the relationship between these astrophysical parameters.

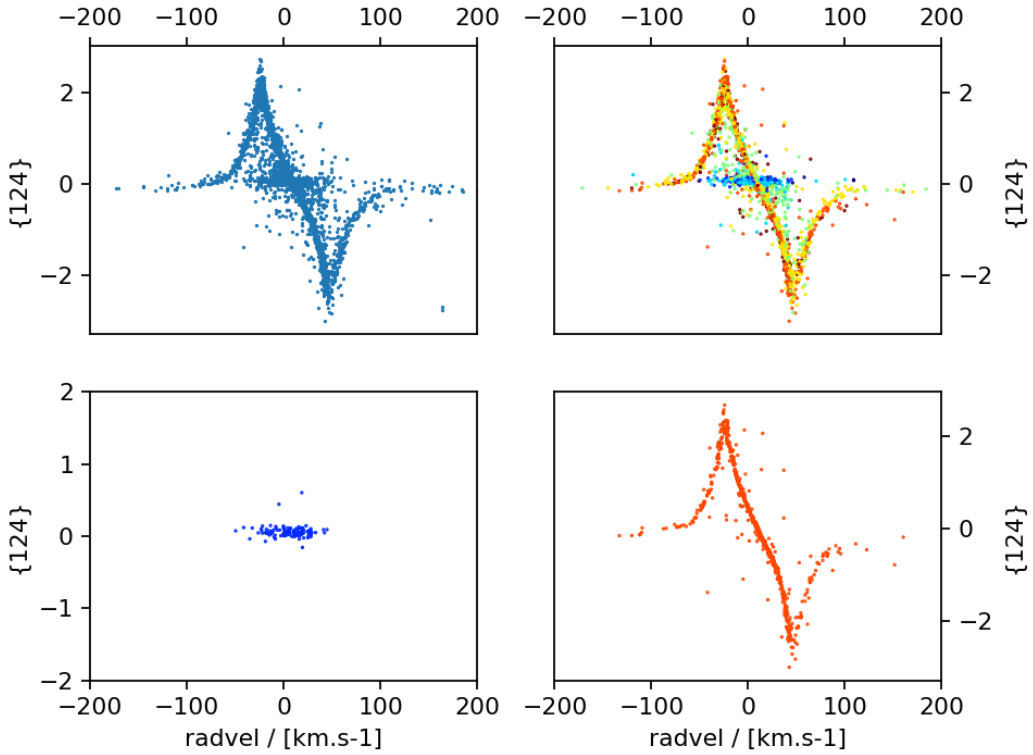


Figure 7.6: Correlation between node 124 values and radial velocity. The top-left panel shows all the data points that have radial velocity labels to compare node values to. Data points that have a spectral type are colour-coded in the top-right as in Fig. 7.5. The bottom row shows how different temperatures appear to have been treated differently (left: O-type; right: K-type).

of astrophysics is not obvious, shown by obscure relationships between known radial velocity and effective temperature values and the code. However, there are clear correlations, particularly noticeable when colouring the data based on the different objects' spectral types, indicating that the network has managed to infer real physics completely unsupervised.

While the other informative dimensions do not correlate with known physical values, this does not mean that they do not include physical information. The fact that the network has found information in the data that distinguishes between different spectra, yet appears ambiguous to us, suggests that there are underlying physical relations. This may include currently unknown astrophysics that may be worth pursuing in the future.

7.4 Telluric rejection

While training our models, we noticed interesting behaviour in the reconstruction of spectra that contain prominent telluric lines. The network treats these lines completely differently to stellar lines, omitting them from reconstruction, such as in the example shown in Fig. 7.7. Autoencoders are trained to reconstruct data via compression at the latent space bottleneck. This means that it needs to preserve what it deems useful information, hence, telluric lines do not appear to be 'useful'. The data that is run through the network is in barycentric coordinates, in which spectra are transformed into a space as though they are observed from the centre of mass between the Earth and Sun (to account for the Earth's motion with respect to the Sun)[1]. In this coordinate system, the position of stellar absorption lines depends only on the individual star's relative velocity. However, from this perspective, there is no clear relation between the spectral continuum and position of telluric lines, and the network is likely rejecting them as noise as they can't be used as information to improve reconstruction. Trained on many spectra, random-distributed noise is already rejected from the reconstructions, as shown previously in Fig. 7.4.

[1]The alternative is topocentric, an Earth-centric system which is how the spectra were originally observed. Telluric lines appear at consistent wavelengths in this system.

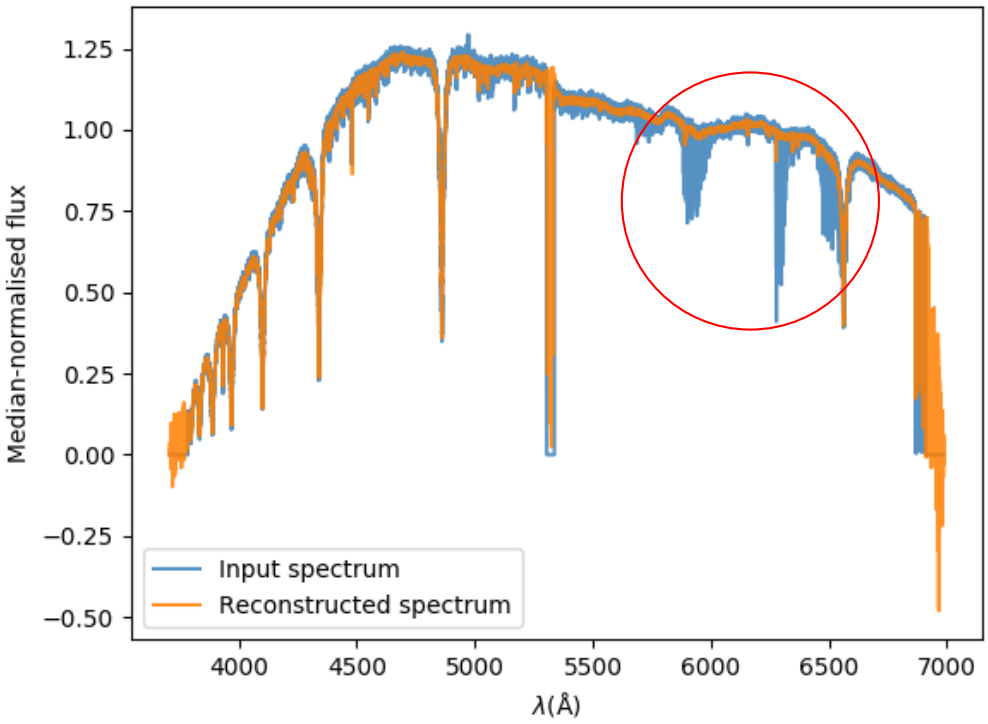


Figure 7.7: Blue is the input spectrum and orange is the reconstructed version. The autoencoder can decompose physically-meaningful components (telluric lines) out of the input. The region annotated by the red circle indicates a high density of such reconstruction rejections. Noise in the reconstructed spectrum is reduced compared to the input, while stellar absorption features are preserved. The reconstruction in the zero-regions is somewhat chaotic as the loss function is only computed for non-zero-regions.

Further work is needed to quantify the quality of telluric rejection, as this aspect of the network's output could be a useful tool. This would require an appropriate metric that incorporates a pseudo-truth, such as MOLECFIT[2], a tool that corrects for atmospheric absorption independent from the autoencoder.

[2] http://www.eso.org/sci/software/pipelines/skytools/molecfit

Chapter 8

Conclusions

In this thesis we have presented research using AI methods in two different astronomical contexts. In the main project we find that despite limitations due to survey constraints such as spectroscopic resources, machine learning and data science techniques can be successfully used to increase the size of a classified sample of Type Ia supernovae. This in turn is expected to enable constraining of cosmological parameters to an unprecedented degree of precision. In a separate project, we find that machines can infer real astrophysical parameters from a network's own interpretation of data using deep learning. These are two specific examples of applying AI, taking very different approaches to different tasks.

8.1 Optimising a magnitude-limited training sample of supernovae

4MOST-TiDES expects to obtain the largest spectroscopically confirmed sample of supernovae to date (>30,000), including Type Ia supernovae which will be used for precision cosmology. However, the transients that are not followed up spectroscopically may still be useful for cosmology. Herein lies the necessity for photometric classification. Using the capabilities and survey constraints of 4MOST, we forecast a spectroscopic sample of supernovae that is magnitude-

limited, reaching $r_{\rm AB} \approx 22.5$ mag. Using machine-learning algorithms, we find the greatest success in the results of photometric classification when we combine this sample with fainter supernovae obtained from larger spectroscopic facilities and then augment the whole sample, to be used as a training set. Whilst on its own, 4MOST cannot give us a fully representative training sample, the accumulated dataset will provide an important basis for a training sample to photometrically classify other LSST transients for which we have host-galaxy redshifts. Including our photometrically classified sample, we expect to produce the largest ever cosmological sample of Type Ia supernovae by more than an order of magnitude.

We started by demonstrating that a representative training sample (of size 1,103) will yield good classification results with SNMACHINE through supervised learning. Algorithms achieve AUC > 0.9 and consistently high purities reaching 95% (with the exception of ANN, although it is important to note that ANN will outperform the other algorithms with much larger training samples). This success is attributed to the fact that the algorithms are trained on features associated with the full range of magnitudes and redshifts in the test set. However, we find that a representative training sample of this nature will not be easily attainable with present spectroscopic facilities. These tests using representative training were also carried out to investigate the role of redshift as an additional feature for classification. We find a consistent improvement in AUC scores when including redshift, demonstrated by a noticeable increase in mean and median over 20 runs. Our results are similar to those in Lochner et al. (2016), although we consider inclusion of redshift important due to its significant impact on classification performance, in contrast to their conclusion that redshift is a relatively unimportant feature. Going from no redshift to photometric and spectroscopic redshifts respectively, we find an increase in average AUC over 20 runs from 0.959 to 0.962 and 0.964 for KNN, 0.947 to 0.963 (both redshifts) for SVM, 0.914 to 0.934 (both redshifts) for ANN and 0.953 to 0.965 and 0.966 for BDT. There appears to be no clear winner between photometric or spectroscopic redshift for this particular simulated dataset, both achieving very similar results. This is surprising, given the fact that photometric redshifts are usually less accurate and less precise than spectroscopic redshifts. We attribute the result to the minimal scatter between spectroscopic and photometric redshifts in the SPCC

simulations; the root mean squared error in photometric redshifts is very small (0.028). However, we find that when the training sample is magnitude-limited, it is less clear whether having redshift helps in the training process or not.

Unfortunately, the SPCC dataset is not large enough that we can fully simulate a 4MOST spectroscopic sample. We are only simulating approximately 1.6% of the full TiDES sample. We find that when considering a spectroscopic sample that is magnitude-limited based on our success criteria and considering 4MOST's capabilities, there are so few objects in the SPCC (approximately 500 after scaling down by a factor of 2, out of 21,319 in total, as discussed in Section 4.4.4) that our results are sensitive to specific choices of which supernovae we include in our training. Despite the variation and spread of results, it is clear that a magnitude limit implies a non-representative training sample that has poor coverage of the test-set feature-space, and, therefore, very negatively affects our results. This does mean, however, that any significant improvement to the performance of the SNMACHINE algorithms when dealing with magnitude-limited training samples is promising. The full TiDES sample size may improve the performance of the magnitude-limited training somewhat, but it will still suffer from the lack of coverage at faint magnitudes and high redshifts. Following this investigation we simulated a full TiDES sample (∼30,000 objects) in our contamination study using a LSST-specific dataset with the software SUPERNNOVA (see § 8.2).

With our 4MOST magnitude-limited training sample as a basis, we next investigated how our results change when combining with additional faint supernovae. A realistic scenario for following up LSST alongside 4MOST would be obtaining spectra of fainter supernovae using facilities such as the VLT and ELT. We simulate such a scenario with the dataset, extending the training to high redshift and increasing the sample size by ∼1,000 supernovae. Over 10 runs we see an increase in the average AUC from 0.547 to 0.961 for KNN, 0.671 to 0.960 for SVM, 0.702 to 0.946 for ANN and 0.628 to 0.969 for BDT. In particular, both KNN and BDT achieved a classified sample of 95% purity in all 10 runs. This is a substantial boost to our results from our orginal training sample, although, on its own, it is not the most successful that we tested. Our results show that complementary faint objects can significantly improve upon a 4MOST magnitude-limited training sample.

We next consider data augmentation to investigate further improvement. By creating artificial light curves, the size is limited only computationally, although we find that results plateau around an augmentation factor of 40–50 new objects per original supernova. Applying AVOCADO to the SPCC, we increase training sample size by a factor of 50. For the augmented magnitude-limited sample we reach average AUC scores of 0.955 for KNN, 0.942 for SVM, 0.965 for ANN and 0.955 for BDT – a large increase, but, with the exception of ANN, not as successful as our combined magnitude-limited and faint training sample. When we augment the combined magnitude-limited and faint sample, we achieve our best AUC scores. However, there is a slight dependence on whether we train our machine learning algorithms to recognise supernovae as either Type Ia or non-Ia, or Type Ia, Ibc or II. The highest average AUCs are 0.985 for KNN, 0.982 for SVM, 0.986 for ANN and 0.984 for BDT and all algorithms are able to reach 95% purity in all 10 runs for this training sample. Considering three classes appears most beneficial for SVM, as this is the only type of training we tested that resulted in all 10 runs reaching 95% purity for this algorithm. For BDT, two classes is more favourable, and for KNN and ANN there is little to no difference. We attribute the success in classification (consistently high AUC and purity) to the fact that including fainter supernovae adds some real constraints to the wavebands at faint magnitudes, i.e. AVOCADO does not need to extrapolate from a set of bright, low-redshift supernovae, as it did when augmenting a purely magnitude-limited sample.

TiDES plans to target every possible transient that is brighter than $r_{\mathrm{AB}} = 22.5$ mag. In this work we assume that we have the full 4MOST-TiDES spectroscopic sample as a training data basis. Hence, our focus on optimisation is how to improve classification using this sample. However, there may be room for further optimisation in survey strategy in how we decide which transients to target that are just below this magnitude limit. Initially, we consider classification when using a hypothetical representative sample, although this may not reflect a fully optimised training sample. A fully optimised sample may require relatively overpopulated bins at high and low redshifts when compared to a 'representative' sample. Achieving this in a spectroscopic follow-up survey would likely need to make use of active learning, such that we observe objects that give the most

improvement to the classification algorithm, following such methods as those presented in Ishida et al. (2019).

Starting with a magnitude-limited training sample constrained by the capabilities of 4MOST, we find that it is optimised when combined with complementary faint supernovae and then augmented to have more coverage of the corresponding test set feature-space. Augmentation is a necessary step to create the most successful realistic training samples, although in future work it will be necessary to test how cosmological assumptions for augmentation could be creating potential bias. Furthermore, in our simulations we assume that the classifications in our spectroscopic sample are 100% correct. Hence, we next tested whether misclassification of 4MOST spectra could have a big impact on photometric classification that could in turn propagate through to the resultant cosmology. Hence, we would want to investigate whether mis-classification of a 4MOST spectrum could propagate through the machine-learning pipeline and affect results, and ultimately the resultant cosmology we determine using our classified sample. These tests would ideally be done with a much larger dataset of supernovae to better reflect what we can do in reality.

8.2 Effects of contaminated training

Following our tests on optimising training, we studied the effects of spectroscopic misclassification in a training sample using a much larger simulated dataset. In a divergence from feature engineering and machine learning we classify transients using a deep recurrent neural network from the software SuperNNova.

From a very large simulated dataset of LSST transients, we take a subset to focus on a TiDES sample of ∼30,000 spectroscopically classified objects. This is randomly sampled from the dataset and is therefore representative of corresponding test data. We take this approach as SuperNNova does not have infrastructure to choose a specific selection of objects, e.g. based on magnitude. Due to time constraints and disruption caused by the COVID-19 pandemic, we did not create a method to customise a training sample as we did for snmachine, but rather

chose a seed to test on a consistent selection of objects for training in order to study effects of contamination.

We first conducted many tests studying how different hyperparameters affect photometric classification in this dataset. Training is incredibly efficient when using cyclic learning, in which the network's learning rate is varied. Running on a GPU, the model converges in less than 10 minutes, overfitting much less than for the case without cyclic learning. We train the model over 30 epochs, include spectroscopic redshift information and apply a cosmology normalisation (normalising each light curve to the maximum flux in any filter) in order to remove any bias learnt by the network matching flux information of Type Ia supernovae with redshift. Further optimisation of hyperparameters may be possible, although we find high success with this combination and keep them consistent throughout our contamination study.

Using these hyperparameters and a perfect, uncontaminated training sample, we run SuperNNova's classification pipeline on a fixed seed to keep the same selection of objects. We observe variation due to probabilistic computations in training the model. Contrary to our primary investigation, we can also consider accuracy as a useful metric as SuperNNova enforces class balance between objects of Type Ia and non-Ia. Over 10 runs we achieve an average: AUC of 0.9845 (on par with our best performing training sample in snmachine); purity of 93.65%; accuracy of 94.08%; efficiency of 94.73%.

We apply contamination by changing the class labels of training and validation objects following predicted error rates presented in Muthukrishna et al. (2019b), ensuring that the corresponding test sample is kept consistent. This assumes that we use DASH to classify TiDES transient spectra. Despite expecting to see a significant drop in classification performance, the spread of results is very similar to those for an uncontaminated training sample. Even if we increase contamination to 5% we find that there is little impact on the photometrically classified sample, in which purity and accuracy only drop by only 1% and 0.1% respectively. AUC stays the most consistent, only varying at the fourth significant figure. We observe a slight increase in efficiency, however this does not mean that classification performance improves. Combined with purity, this implies that

more objects of both Type Ia and non-Ia are classified as Ia, due to increased uncertainty in how to classify test objects.

To take this study further, we would require taking our analysis through to constraining cosmology with our photometric sample. This would help us understand how contamination propagates through to measurements of parameters such as the dark energy equation of state. As Vincenzi et al. (2021a) found that errors in photometric classification do not introduce significant bias in such measurements, these results suggest that the effects of contamination in both spectroscopic training and photometrically classified samples will not greatly impact cosmology.

8.3 Deep learning of stellar spectra

In a departure from supernova classification, we present the work undertaken during a data science internship studying stellar spectra with deep learning. The ESO archive consists of a huge range of astronomical objects. This includes observations from the HARPS instrument, which is dedicated to the discovery of exoplanets. Using the ∼270,000 spectra from the HARPS dataset we apply a data-driven approach to classification with unsupervised learning. Through a reconstruction task in an autoencoder neural network, we create a low-dimensional representation of the data at the information bottleneck.

We determine a sample of 7,653 unique HARPS objects based on the names from the dataset metadata. We make an effort to identify unique objects by correcting for string format, however some contamination is unavoidable without applying substantial human input, which we want to avoid. Clustering methods are insufficient in determining unique objects as the metadata coordinates and number of observations per object vary too much to produce a robust sample of unique objects. Identifying unique objects by name instead of clustering also removes added contamination of solar system objects that may appear in very different parts of the sky. We use the unique sample so that we can study what the network has learned with respect to each object from the latent representation.

Different dimensions of latent space are tested and we consider both a deterministic autoencoder, and a probabilistic variational autoencoder. The network model is developed by reconstruction training through minimisation of the loss function. Good reconstruction implies that the network has been able to cipher important information in the data. Reconstruction of spectra is better in the deterministic network, although our main focus is on what the network learns in the latent represenation, rather than achieving perfect reconstruction. In the variational autoencoder we can apply disentanglement to enforce orthogonality between latent dimensions. This creates a network that assigns different nodes to different information that is encoded in the spectra. More latent dimensions implies better reconstruction, although interestingly we find that both 8 and 128 dimensions yield six informative dimensions, defined by the median absolute deviation of statistical dispersion.

Two of these informative dimensions show clear correlations with known physical labels. By compressing the data into a low-dimensional representation, the network has learned notions of effective temperature and radial velocity. The relationships are reinforced when identifying spectral types of objects. These dimensions describe some of the network's own classification of the HARPS spectra, an interpretation that shows unsupervised learning of real astrophysics. The other dimensions' interpretation appears ambiguous although leaves significant room for further study. To take this work further we would also like to generalise a model to the whole of the ESO archive, to analyse spectra from an even wider range of astronomical objects.

8.4 Summary

We applied supervised learning using supernova class labels in preparation for the next generation of cosmological measurements. TiDES will conduct a spectroscopic follow-up campaign of LSST transients, although is magnitude-limited due to the capabilities of 4MOST. To utilise machine-learning algorithms effectively, we require multiple spectroscopic surveys from different facilities, in combination with data augmentation. Following these procedures we can construct a training sample

that is much more representative and achieve a very high level of purity in our photometrically classified sample. Following these approaches will considerably increase the Type Ia supernova sample from LSST and hence greatly improve constraints on cosmological parameters. Measurements of the density of matter Ω_m and the dark energy equation of state w will progress our physical understanding of the Universe, including its content, history and fate.

Using a deep recurrent neural network, we classify supernovae with a simulated representative training sample. Contamination of the training sample following expected error rates from the literature appears to have minimal effect on classification performance. Combined with the little impact on cosmology from errors in photometric classification, this offers a promising outlook for cosmological measurements in upcoming surveys.

Finally, we delve into the vast ESO archive in an unsupervised, data-driven approach to classification of stellar spectra. By training an autoencoder network to reconstruct spectra, it learns a low-dimensional representation of the data. Through disentanglement, the network produces six informative dimensions that describe the data. We find that two of these contain notions of real astrophysics, illustrated by correlation between latent node values with effective temperature and radial velocity labels. Interpretation of the other dimensions appears ambiguous, but, with further study, may reveal new patterns characterising other physics hidden in the data.

We have demonstrated original research using multiple artificial intelligence and data science techniques in an astronomical context. In conclusion, AI is vital to make the most of available data in present and future research. The power of machine-learning and deep-learning algorithms has been demonstrated: in a deviation from traditional classification methods using pure spectroscopy, supernova light curves can successfully be classified with an optimised realistic training sample to aid progress in our understanding of the Universe; with minimal human input and 'letting the data speak for itself', machines can determine their own interpretation of different stars and infer real astrophysical parameters from stellar spectra.

References

Abbott B. P., et al., 2016, Phys. Rev. Lett., 116, 061102

Abbott T. M. C., et al., 2019, ApJL, 872, L30

Aizerman M. A., Braverman E. A., Rozonoer L., 1964. No. 25 in Automation and Remote Control. pp 821–837

Alam S., et al., 2017, MNRAS, 470, 2617

Altman N. S., 1992, Am. Stat., 46, 175

Amendola L., Tsujikawa S., 2010, Dark Energy: Theory and Observations. Cambridge University Press, doi:10.1017/CBO9780511750823

Balland C., et al., 2009, A&A, 507, 85

Baron D., 2019, arXiv:1904.07248 [astro-ph.IM]

Beck R., Lin C. A., Ishida E. E. O., Gieseke F., de Souza R. S., Costa-Duarte M. V., Hattab M. W., Krone-Martins A., 2017, MNRAS, 468, 4323

Bengio Y., Courville A., Vincent P., 2013, IEEE PAMI, 35, 1798

Bennett C. L., et al., 2013, ApJS, 208, 20

Betoule M., et al., 2014, A&A, 568, A22

Blake C., Glazebrook K., 2003, ApJ, 594, 665

Blum A. L., Langley P., 1997, Artif. Intell., 97, 245

Blundell C., Cornebise J., Kavukcuoglu K., Wierstra D., 2015, arXiv:1505.05424 [stat.ML]

Boffin H. M. J., Jerabkova T., Mérand A., Stoehr F., 2019, The Messenger, 178, 61

Boggess N. W., et al., 1992, ApJ, 397, 420

Boone K., 2019, AJ, 158, 257

Brzeski J., et al., 2018, in Evans C. J., Simard L., Takami H., eds, Vol. 10702, Ground-based and Airborne Instrumentation for Astronomy VII. SPIE, pp 2172 – 2188, doi:10.1117/12.2310062

Burgess C. P., Higgins I., Pal A., Matthey L., Watters N., Desjardins G., Lerchner A., 2018, arXiv:1804.03599 [stat.ML]

Caldwell R. R., Dave R., Steinhardt P. J., 1998, Phys. Rev. Lett., 80, 1582

Camarena D., Marra V., 2020, Phys. Rev. Research, 2, 013028

Carrick J. E., Hook I. M., Swann E., Boone K., Frohmaier C., Kim A. G., Sullivan M., 2021, MNRAS, 508, 1

Carroll B. W., Ostlie D. A., 2007, An Introduction to Modern Astrophysics, 2nd (international) edn. Cambridge University Press

Chandrasekhar S., 1931, ApJ, 74, 81

Charnock T., Moss A., 2017, ApJL, 837, L28

Chawla N. V., Japkowicz N., Kotcz A., 2004, SIGKDD Explor. Newsl., 6, 1–6

Chetlur S., Woolley C., Vandermersch P., Cohen J., Tran J., Catanzaro B., Shelhamer E., 2014, arXiv:1410.0759 [cs.NE]

Ciliegi P., et al., 2021, The Messenger, 182, 13

Cohn D., Ghahramani Z., Jordan M., 1996, J. Artif. Intell. Res., 4, 129

Cortes C., Vapnik V., 1995, Mach. Learn., 20, 273

de Bernardis P., et al., 2000, Nature, 404, 955–959

de Jong R. S., et al., 2019, The Messenger, 175, 3

Doggett J. B., Branch D., 1985, AJ, 90, 2303

du Plessis M. C., Sugiyama M., 2014, Neural Networks, 50, 110

Dyson F. W., Eddington A. S., Davidson C., 1920, Philos. Trans. R. Soc. Series A, (Containing Papers of a Mathematical or Physical Character), 220, 291

Einasto J., 2001, arXiv:astro-ph/0012161

Eisenstein D. J., et al., 2005, ApJ, 633, 560

Event Horizon Telescope Collaboration et al., 2019, ApJL, 875, L1

Filippenko A. V., 1997, Annu. Rev. Astron. Astrophys., 35, 309

Filippenko A. V., et al., 1992, Annu. Rev. Astron. Astrophys., 104, 1543

Fixsen D. J., 2009, ApJ, 707, 916

Fixsen D. J., Cheng E. S., Gales J. M., Mather J. C., Shafer R. A., Wright E. L., 1996, ApJ, 473, 576

Fluke C. J., Jacobs C., 2020, WIREs Data Mining and Knowledge Discovery, 10, e1349

Foley R. J., et al., 2013, ApJ, 767, 57

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

Friedman J. H., 2002, Comput. Stat. Data Anal., 38, 367–378

Frohmaier C., et al., 2019, MNRAS, 486, 2308

Gal Y., Ghahramani Z., 2016, arXiv:1506.02142 [stat.ML]

Goodfellow I., Bengio Y., Courville A., 2016, Deep Learning. MIT Press

Graham M. L., et al., 2020, AJ, 159, 258

Guth A. H., 1981, Phys. Rev. D, 23, 347

Guy J., et al., 2007, A&A, 466, 11

Guy J., et al., 2010, A&A, 523, A7

Géron A., 2017, Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. O'Reilly Media, Sebastopol, CA

Hamuy M., Phillips M. M., Suntzeff N. B., Schommer R. A., Maza J., Aviles R., 1996, AJ, 112, 2391

Hansen B. M. S., et al., 2002, ApJ, 574, L155

Hayden B., et al., 2021, ApJ, 912, 87

Heger A., Fryer C. L., Woosley S. E., Langer N., Hartmann D. H., 2003, ApJ, 591, 288

Heisenberg W., 1958, Physics and Philosophy: The Revolution in Modern Science. Harper: New York

Hinton G., Sejnowski T. J., 1999, Unsupervised Learning: Foundations of Neural Computation. The MIT Press, doi:10.7551/mitpress/7011.001.0001

Hložek R., et al., 2020, arXiv:2012.12392 [astro-ph.IM]

Hogan C. J., Kaiser N., Rees M. J., Fabbri R., McCrea W. H., Lynden-Bell D., 1982, Philos. Trans. R. Soc., 307, 97

Hogg D. W., Baldry I. K., Blanton M. R., Eisenstein D. J., 2002, arXiv:astro-ph/0210394

Hook I. M., 2013, Philos. Trans. R. Soc., 371, 20120282

Hotelling H., 1933, J. Educ. Psychol., 24, 417

Hsiao E. Y., Conley A., Howell D. A., Sullivan M., Pritchet C. J., Carlberg R. G., Nugent P. E., Phillips M. M., 2007, ApJ, 663, 1187

Hubble E., 1929, PNAS, 15, 168

Huterer D., Shafer D. L., 2017, Rep. Prog. Phys., 81, 016901

Huterer D., Turner M. S., 1999, Phys. Rev. D, 60, 081301

Ishida E. E. O., et al., 2019, MNRAS, 483, 2

Ivezić Ž., et al., 2019, ApJ, 873, 111

Jassal H. K., Bagla J. S., Padmanabhan T., 2005, MNRAS, 356, L11

Jones B. J., Wyse R. F., 1985, A&A, 149, 144

Jones D. O., et al., 2017, ApJ, 843, 6

Kessler R., et al., 2009, Public. Astron. Soc. Pac., 121, 1028

Kessler R., Conley A., Jha S., Kuhlmann S., 2010a, arXiv:1001.5210 [astro-ph.IM]

Kessler R., et al., 2010b, Public. of the Astron. Soc. Pac., 122, 1415

Kessler R., et al., 2019, Public. of the Astron. Soc. Pac., 131, 094501

Kingma D. P., Welling M., 2014, Auto-Encoding Variational Bayes

Knop R. A., et al., 2003, ApJ, 598, 102

Kohavi R., 1995, in Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2. IJCAI'95. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 1137–1143

Kullback S., Leibler R. A., 1951, Ann. Math. Statist., 22, 79

LSST Science Collaboration et al., 2009, arXiv:0912.0201 [astro-ph.IM]

LSST Science Collaboration et al., 2017, arXiv:1708.04058 [astro-ph.IM]

Laureijs R., et al., 2011, arXiv:1110.3193 [astro-ph.CO]

Li L.-X., 2008, MNRAS, 388, 1487

Li W., et al., 2003, Public. Astron. Soc. Pac., 115, 453

Li W., et al., 2011, MNRAS, 412, 1441

Li J., Cheng K., Wang S., Morstatter F., Trevino R. P., Tang J., Liu H., 2017, ACM Comput. Surv., 50

Licata G., 2015, Journal of Computer Science & Systems Biology, 8, 124

Liddle A., 2003, An Introduction to Modern Cosmology; 2nd ed. Wiley

Liddle A. R., Lyth D. H., 2000, Cosmological Inflation and Large-Scale Structure. Cambridge University Press

Linder E. V., 2003a, Phys. Rev. D, 68, 083504

Linder E. V., 2003b, Phys. Rev. Lett., 90, 091301

Linder E. V., 2021, arXiv:2106.09581 [astro-ph.CO]

Linder E. V., Mitra A., 2019, Phys. Rev. D, 100, 043542

Lochner M., McEwen J. D., Peiris H. V., Lahav O., Winter M. K., 2016, ApJS, 225, 31

Lochner M., et al., 2018, arXiv:1812.00515 [astro-ph.IM]

Lochner M., et al., 2021, arXiv:2104.05676 [astro-ph.CO]

Maas A. L., Hannun A. Y., Ng A. Y., 2013, in ICML Workshop on Deep Learning for Audio, Speech and Language Processing.

MacKay D. J. C., 2003, Information Theory, Inference, and Learning Algorithms. Cambridge University Press

Mehri S., Kumar K., Gulrajani I., Kumar R., Jain S., Sotelo J., Courville A., Bengio Y., 2017, arxiv:1612.07837 [cs.SD]

Mickaelian A. M., 2020, ComBAO, 67, 159

Mitra A., Linder E. V., 2021, Phys. Rev. D, 103, 023524

Möller A., de Boissière T., 2020, MNRAS, 491, 4277

Murakami Y. S., Stahl B. E., Zhang K. D., Chu M. R., McGinness E. C., Patra K. C., Filippenko A. V., 2021, Monthly Notices of the Royal Astronomical Society: Letters, 504, L34

Muthukrishna D., Narayan G., Mandel K. S., Biswas R., Hložek R., 2019a, Public. Astron. Soc. Pac., 131, 118002

Muthukrishna D., Parkinson D., Tucker B. E., 2019b, ApJ, 885, 85

Napiwotzki R., Schoenberner D., Wenske V., 1993, A&A, 268, 653

Oizumi M., Albantakis L., Tononi G., 2014, PLOS Computational Biology, 10, 1

Oke J. B., Sandage A., 1968, ApJ, 154, 21

Owens J. D., Houston M., Luebke D., Green S., Stone J. E., Phillips J. C., 2008, Proceedings of the IEEE, 96, 879

Pasquet J., Pasquet J., Chaumont M., Fouchez D., 2019, A&A, 627, A21

Pearson K., 1901, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2, 559

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Penrose R., 1989, The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics. Oxford University Press, Inc., USA

Pepe F., et al., 2002, The Messenger, 110, 9

Percival W. J., et al., 2010, MNRAS, 401, 2148

Perlmutter S., et al., 1999, ApJ, 517, 565

Perrett K., et al., 2012, AJ, 144, 59

Phillips M. M., 1993, ApJL, 413, L105

Planck Collaboration et al., 2020, A&A, 641, A6

Platt J. C., 1999, in Advances in Large Margin Classifiers. MIT Press, pp 61–74

Qu H., Sako M., Möller A., Doux C., 2021, AJ, 162, 67

Ratra B., Peebles P. J. E., 1988, Phys. Rev. D, 37, 3406

Revsbech E. A., Trotta R., van Dyk D. A., 2018, MNRAS, 473, 3969

Richardson D., Jenkins Robert L. I., Wright J., Maddox L., 2014, AJ, 147, 118

Riess A. G., et al., 1998, AJ, 116, 1009

Riess A. G., et al., 2018, ApJ, 853, 126

Rosenblatt F., 1960, Proceedings of the IRE, 48, 301

Rubin V. C., Ford W. Kent J., 1970, ApJ, 159, 379

Rumelhart D. E., Hinton G. E., Williams R. J., 1986, Nature, 323, 533

Salamanca I., Terlevich R. J., Tenorio-Tagle G., 2002, MNRAS, 330, 844

Scolnic D. M., et al., 2018a, arXiv:1812.00516 [astro-ph.IM]

Scolnic D. M., et al., 2018b, ApJ, 859, 101

Scuderi S., Bonanno G., Spadaro D., Panagia N., Lamers H. J. G. L. M., de Koter A., 1994, ApJ, 437, 465

Sedaghat N., Romaniello M., Carrick J. E., Pineau F.-X., 2021, MNRAS, 501, 6026

Seikel M., Clarkson C., Smith M., 2012, J. Cosmol. Astropart. Phys., 2012, 036

Simonyan K., Zisserman A., 2015, arXiv:1409.1556 [cs.CV]

Slipher V., 1912, Lowell Observatory Bulletin, 2, 56

Smith L. N., 2017, arXiv:1506.01186 [cs.CV]

Smith M., et al., 2020, AJ, 160, 267

Stassun K. G., et al., 2019, AJ, 158, 138

Sutskever I., Vinyals O., Le Q. V., 2014, arXiv:1409.3215 [cs.CL]

Swann E., et al., 2019, The Messenger, 175, 58

Taubenberger S., 2017, The Extremes of Thermonuclear Supernovae. Springer International Publishing, Cham, pp 317–373, doi:10.1007/978-3-319-21846-5˙37

Taylor M., et al., 2014, ApJ, 792, 135

Tegmark M., 2017, Life 3.0: Being Human in the Age of Artificial Intelligence. Knopf Publishing Group

Tempel E., et al., 2020a, MNRAS, 497, 4626

Tempel E., et al., 2020b, A&A, 635, A101

The LSST Dark Energy Science Collaboration et al., 2018, arXiv:1809.01669 [astro-ph.CO]

Tishby N., Pereira F. C., Bialek W., 1999, in Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing. pp 368–377

Tononi G., 2004, BMC Neuroscience, 5, 42

Tripp R., 1998, A&A, 331, 815

Tschannen M., Bachem O., Lucic M., 2018, arXiv:1812.05069 [cs.LG]

Van der Maaten L., Hinton G., 2008, JMLR, 9, 2579

Vincenzi M., Sullivan M., Firth R. E., Gutiérrez C. P., Frohmaier C., Smith M., Angus C., Nichol R. C., 2019, MNRAS, 489, 5802

Vincenzi M., et al., 2021a, arXiv:2111.10382 [astro-ph.CO]

Vincenzi M., et al., 2021b, MNRAS, 505, 2819

Walker E. S., et al., 2010, MNRAS, 410, 1262

Weiss G., Provost F., 2003, J. Artif. Intell. Res. (JAIR), 19, 315

Wenger, M. et al., 2000, Astron. Astrophys. Suppl. Ser., 143, 9

Werbos P. J., 1974, PhD thesis, Harvard University

White M., Scott D., Silk J., 1994, Annual Review of Astronomy and Astrophysics, 32, 319

Williams S. C., et al., 2020, MNRAS, 495, 3859

Winkler R., Haynes D. M., Bellido-Tirado O., Xu W., Haynes R., 2014, in Angeli G. Z., Dierickx P., eds, Vol. 9150, Modeling, Systems Engineering, and Project Management for Astronomy VI. SPIE, pp 288 – 297, doi:10.1117/12.2056463

Zwicky F., 1937, ApJ, 86, 217