Identification and Classification of Emission-line Stars in the GALAH Survey Using Machine Learning

By

Praveen Nisal Jayasuriya Daluwathumullagamage

A thesis submitted to Macquarie University

for the degree of Master of Research

The School of Mathematical and Physical Sciences

September 2022

MACQUARIE
University
SYDNEY·AUSTRALIA

Examiner's Copy

Typeset in LaTeX 2_ε.

This thesis is an account of research undertaken between July 2021 and June 2022 at The School of Mathematical and Physical Sciences, Macquarie University, Sydney NSW, Australia.

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

Praveen Nisal Jayasuriya Daluwathumullagamage

Acknowledgements

I would like to thank my wife Gazala who supported me and pushed me to do more astronomy and apply to the Master's program at Macquarie University. I would also like to thank my family in Sri Lanka who support everything I do and encouraged me to follow my interests no matter what. I would not be completing this work if not for my family.

A big thank you to my supervisor Dr Daniel Zucker who, in addition to providing invaluable feedback, guided and believed in me every step of the way. Your friendship and mentorship has made this journey so much more enjoyable. Thank you to Dr Ben Montet at UNSW who introduced me to the astronomy community in Australia and encouraged me to apply to Macquarie University. As an immigrant and as someone who had no formal training in astronomy until this point, I owe him a debt of gratitude.

Many thanks to Dr Sarah Martell at UNSW for the early feedback and guidance on P Cygni spectra. Thank you to Dr Gregor Traven at Lund University for his valuable feedback and sanity checks towards the end of the project. Thanks to Arv Hughes, Dr Sven Buder and Dr Klemen Čotar and the various members of the GALAH science team and HDR cohort at Macquarie University who provided feedback and support during various forums and meetings throughout the last year—your kindness, generosity and camaraderie is highly appreciated. Finally, many thanks to all my friends, particularly Nuzhi Meyen in Sri Lanka for his input on the mathematics of machine learning and my many friends in Sydney who have had to endure me talking about astronomy 24/7/365.

Abstract

The advent of massive stellar spectroscopic surveys with hundreds of thousands—or even millions of spectra—presents serious challenges for the identification and classification of atypical objects, such as emission-line stars. To date, a variety of machine-learning methods have been applied, but in most cases actual classification has been carried out manually by humans, even for datasets comprising tens of thousands of spectra. As spectroscopic surveys grow larger—by orders of magnitude—manual solutions become untenable. Additionally, in instances where machine learning has been applied, researchers have relied on manually developed training data sets which are not available for atypical spectra.

This thesis seeks to address the twin problems of identification and classification of emission-line stars in large spectroscopic data sets—like the GALAH survey—through the application of unsupervised machine learning methods. GALAH is a million-star high-resolution spectroscopic survey of the Milky Way, and its most recent public data release (Data Release 3 - DR3) contains more than 600,000 high-resolution spectra.

When developing machine learning methods to identify emission-line spectra in the GALAH survey, some limitations encountered included a lack of training data, a high proportion of typical spectra in survey data—resulting in poor performance of other machine learning methods—and high dimensionality, among other constraints, making the identification and classification of emission-line spectra extremely challenging.

This necessitates the use of unsupervised machine learning methods, which this thesis will demonstrate by identifying and classifying over 7000 emission-line spectra, including over 200 P Cygni and 200 inverse P Cygni spectra, thereby providing a more accurate estimation of the population of emission line stars found in the DR3 survey. This method can, in turn, be applied to other surveys, leading to more emission line stars being identified, while improving

the accuracy of the stellar parameter estimates of these atypical objects.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Modern large scale spectroscopic surveys generate hundreds of thousands, or even millions of spectra (see Table 1.1). The analysis of these high volume data-sets presents a significant challenge to the researcher as well as to compute infrastructure and engineering. Additionally, if the data is collected by a high resolution instrument, the researcher will face the challenge of wrangling and analysing individual data points with dimensions at the scale of several thousands per spectrum (e.g., Buder et al. 2021). When the number of data points (i.e., spectra) is of the order of hundreds of thousands (or millions) and when each data point has a dimensionality of several thousands, it becomes impractical and perhaps even unfeasible to process and analyse these data using manual methods such as naked eye observations of spectral plots.

In the case of stellar spectroscopic surveys, unless the science goals have been set to bias a survey specifically towards, for example, star forming regions (Traven et al. 2015),

these surveys will contain a majority of spectra that are presumably "typical", that is, with properties most common to the types of stars being targeted. Thus the identification and classification of *atypical* objects, such as emission-line stars, presents a serious challenge, in addition to those mentioned previously, as these spectra are outliers or anomalies in an otherwise typical set of stellar spectra.

Survey	**Number of Spectra**	**Resolution**
Gaia ESO	~150,000	R~5000 to R~30,000
LAMOST	~10,000,000	R~500 to R~1800
APOGEE	~250,000	R~22,500
RAVE	~600,000	R~7500
GAIA	~100,000,000	R~11,500 (RVS)

TABLE 1.1: Modern stellar spectroscopic surveys often generate large volumes of data, in some cases at high spectral resolution.

These surveys use data analysis pipelines to derive stellar parameters, and often use template spectra that represent typical or non-peculiar baselines. It has been demonstrated that the use of these so-called non-peculiar baselines can impact the accurate determination of effective temperature (Amarsi et al. 2018; Cayrel et al. 2011; Giribaldi et al. 2019) and stellar mass (Bergemann et al. 2016; Ness et al. 2016), among other key measurements. The identification of atypical emission-line stars can thus help improve the accurate determination of these stellar parameters. To achieve this, once identified and classified, these spectra can be removed from the primary data analysis pipeline containing typical spectra, and can be reduced separately by secondary pipelines more suited for their peculiarities.

The detection of atypical signals or data points in significantly larger, more typical populations of data presents itself well to modern machine-learning methods, particularly to anomaly detection, as well as to clustering methods (unsupervised learning). To date, a variety of machine-learning methods have been applied to the identification of emission-line stars, and in particular to Hα emission-line stars. However, major drawbacks and challenges remain, for both the identification and the classification of emission-line stars, despite the use of popular and seemingly robust machine-learning methods such as dimensionality reduction, k-means clustering and neural networks. Given these challenges, it is not uncommon that

manual methods are still being used for the identification and classification of emission-line stars even in modern data-sets with thousands of stars, with Zhang et al. (2021) being a recent example.

In order to tackle the twin problems of identification and classification of emission-line stars, this work will apply unsupervised machine learning methods to data from the GALAH survey (Buder et al. 2021). The GALAH survey is a million-star high-resolution spectroscopic survey of the Milky Way which uses the HERMES spectrograph at the Anglo Australian Telescope (De Silva et al. 2015). The most recent public data release from GALAH, Data Release 3 (DR3; Buder et al. 2021)), contains more than 600,000 high-resolution spectra. Motivated by the opportunities and challenges presented above, this work presents a novel data-driven approach to identify and classify emission-line stars in the GALAH database, utilising an unsupervised machine learning method that performs spectral-morphology-based clustering, with a particular focus on stars showing P Cygni and inverse P Cygni line Hα profiles. However the methods presented in this work are sufficiently general that they can be applied to other atypical emission-line spectra, the details of which are provided in subsequent chapters. Additionally, considerable attention has been given to the use of computationally efficient algorithms, to ensure that the methods presented scale sufficiently beyond the size and scale of the GALAH survey and can be applied to other spectroscopic surveys where emission-line spectra are present.

1.2 A Note on P Cygni Line Profiles

A principal line profile or feature that the reader will encounter in this work is the P Cygni line profile, which is shown in Figure 1.1. P Cygni (or 34 Cygni) is a luminous blue variable star (LBV) that has been studied extensively (Beals 1953; Elliott et al. 2022; Hutchings 1969; Mizumoto et al. 2018; Underhill 1966). Willem Janszoon Blaeu, a Dutch cartographer and student of the astronomer Tycho Brahe, is considered to have provided the first known set of observations of 34 Cygni in the year 1600 (De Groot & Sterken 2001). The stellar spectrum of 34 Cygni is peculiar. It exhibits the characteristics of a B-type supergiant except that almost all absorption lines are blueshifted with a redshifted emission component (Hutchings 1969). This characteristic line profile can be clearly observed in proximity to the Hα line whose rest wavelength is at $\sim$ 6562.7Å (Traven et al. 2015; Zhang et al. 2021)

P Cygni-type stars or more simply, P Cygni stars, are stars that exhibit line profiles that are similar to the characteristic profile of 34 Cygni. The spectra of these stars show characteristic absorption, emission and wide absorption sub-components (Zhang et al. 2021). The redshifted absorption (and blueshifted emission) counterparts to P Cygni stars have also been observed. These belong to a class of objects known as inverse P Cygni stars.

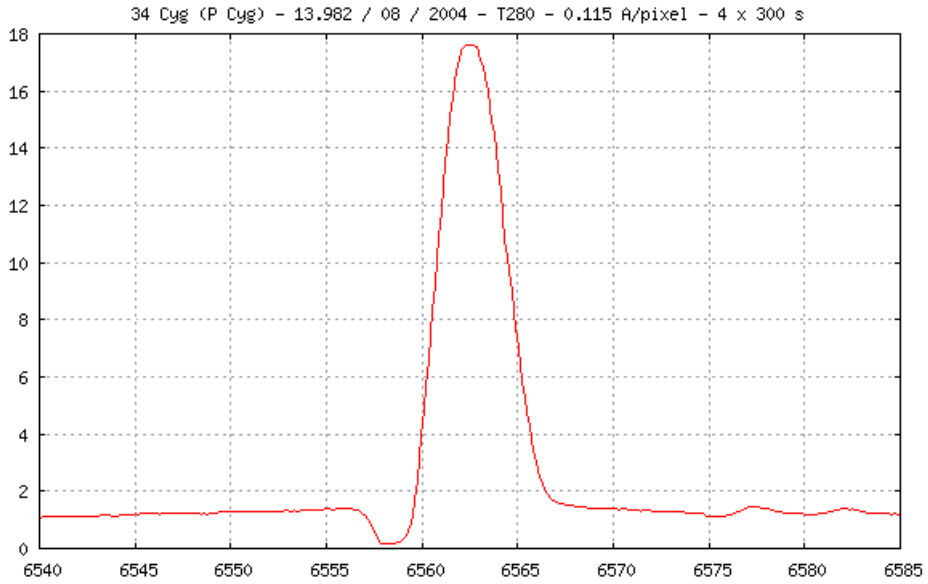

FIGURE 1.1: The normalised spectrum of 34 Cygni around Hα.

It is believed that distinct physical processes within—and around—these stars generate the respective line profiles (Hou et al. 2016). Beals 1953 was the first to demonstrate that P Cygni and inverse P Cygni line profiles can be respectively explained by the expanding—or contracting—shell of gas (e.g., an optically thick wind) surrounding a hot, young, massive star.

An observed P Cygni line profile is thus a result of an expanding shell of gas or optically thick wind around a young hot star. An average or ”normal” main sequence star will only show a deep absorption line/trough near Hα. However, in the case of a P Cygni star, the regions B, C and D (in Figure 1.2) of the shell contribute to an emission line as a result of their excitation by photons from the photosphere of the star. This emission line occurs at a velocity offset from the systemic velocity. As we move from region B to C, the intensity of the emission line increases until we reach the edge of the shell. Beyond this point, the shell is receding with respect to the line of sight and the intensity of the emission line consequently decreases. However, photons emitted by the star towards the observer (A in Figure 1.2) are

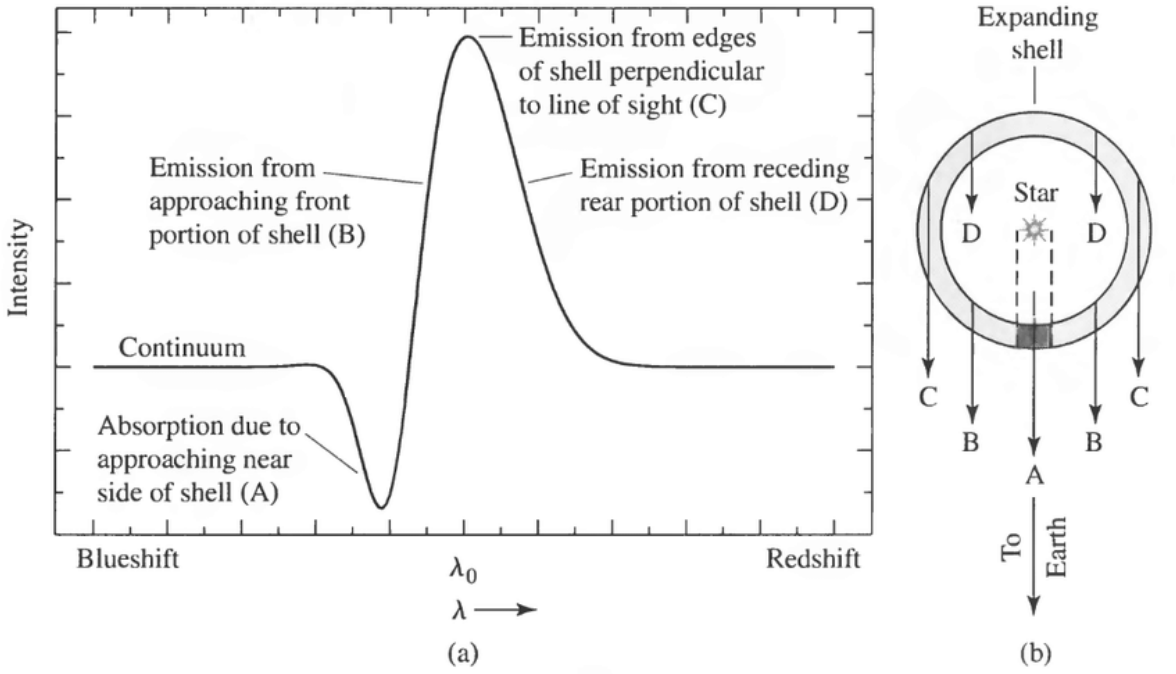

FIGURE 1.2: Cartoon depicting the physical mechanism by which a P Cygni line profile is generated. Reproduced from Kasai (2013).

absorbed and re-emitted in random directions, causing a net drop in flux at a wavelength corresponding to the velocity of the shell. This segment of the shell along the line of sight thus contributes to the generation of the blueshifted absorption line in the spectrum.

It is believed that the opposite process occurs in the case of an inverse P Cygni star. In this case, the shell of gas is contracting, and this inflow is responsible for the blueshifted emission line, often to the blueward of Hα. A full discussion of other classes of emission-line stars such as those presented in Figure 1.3, and the physics that generate the observed line profiles is beyond the scope of this thesis. However, this work will present suitable classes of such candidates in the GALAH survey where relevant, with the key motivation of this work being the identification of P Cygni and inverse P Cygni spectra in a survey that primarily contains typical spectra.

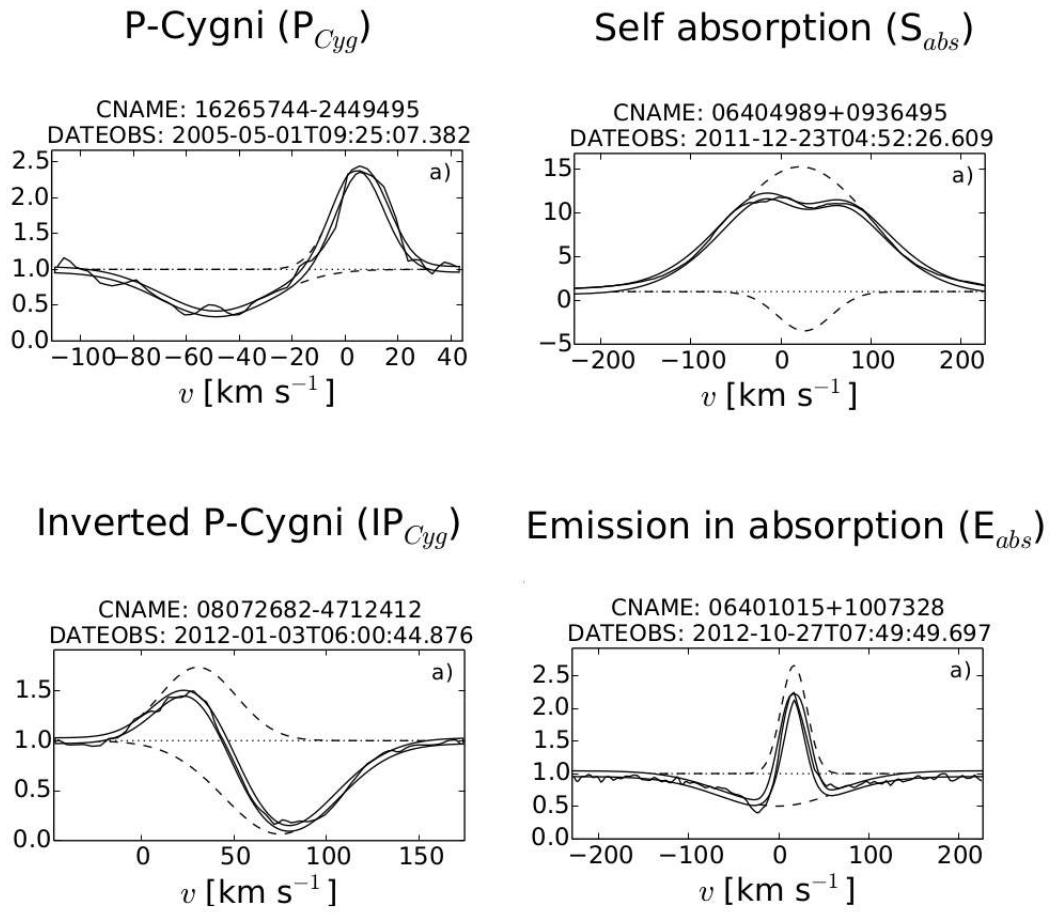

FIGURE 1.3: Four general classes of emission-line spectra identified in the Gaia-ESO Survey. Reproduced from Traven et al. (2015) and edited for clarity.

1.3 A Brief History of Classification

One of the first modern attempts at the identification and classification emission-line stars, in particular P Cygni stars, was by Beals (1953). Beals compiled Northern Hemisphere observations of emission-line stars into a comprehensive catalogue. The catalogue was constructed by examining spectra visually, during a period of observation between the years 1928 and 1946. The data were then used to generate hypotheses of how P Cygni stars may exchange materials with their surroundings via accretion, inflows and outflows, and the morphological properties of the spectra were subsequently measured to calculate the wind velocities of inflows and outflows. Manual classification and visual observation of spectral plots was suitable in this context since the volume of data was not significant. However, such an approach would probably not be feasible in the modern era due to the reasons outlined above.

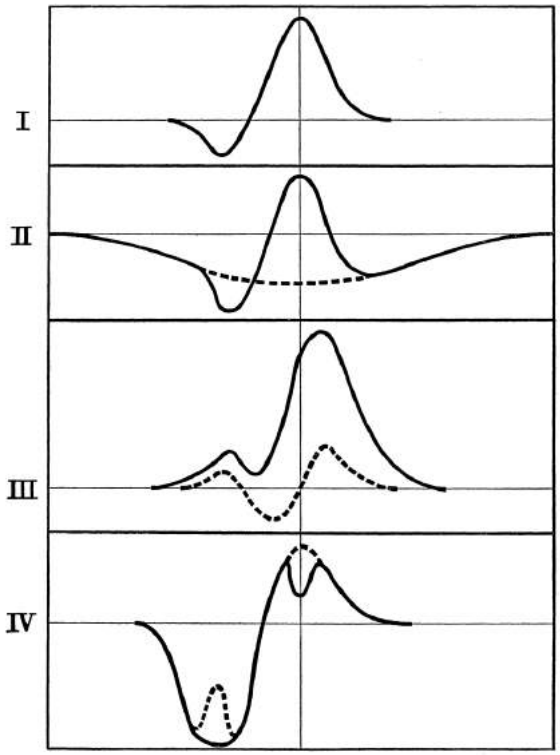

FIGURE 1.4: Primary classes of emission-line spectra proposed by Beals (1953).

Beals' work also presents an early attempt at classifying of emission-line stars based on their morphologies. The classification provided is simpler than more modern schemes, for example those presented in Traven et al. (2015). In addition to what he labelled the primary classes (Figure 1.4), Beals proposed a set of non-typical classes, which were nonetheless also considered to be P Cygni spectra by Beals (Figure 1.5). In this work these non-typical classes are not treated as P Cygni spectra *per se*, but they are considered to be other classes of emission-line spectra. This approach is congruent with modern studies (e.g., Čotar et al. 2021; Reipurth et al. 1996; Zhang et al. 2021) which place constraints on the classification of P Cygni and inverse P Cygni morphologies.

It is worth noting that even subsequent work such as Reipurth et al. (1996) also relied on manual human classification of emission-line stars based on the morphological properties of the spectra. This remained a common practice towards the end of the 20th century, as the volume of data was not sufficiently large to warrant the use of data-driven and machine-learning-based classification routines. With the advent of large scale spectroscopic surveys, however, and significant advances in computational resources and machine learning methods, the stage was finally set for data-driven classification of machine-learning methods such as

the use of t-distributed stochastic neighbour embedding (t-SNE) (Traven et al. 2017) and the application of neural networks as anomaly detectors (Čotar et al. 2021). But as this work will demonstrate, these methods are not sufficiently robust at identifying and classifying P Cygni, inverse P Cygni and other emission-line spectra. In order to overcome these challenges, this work will introduce a novel unsupervised machine-learning approach which will be discussed in following chapters.

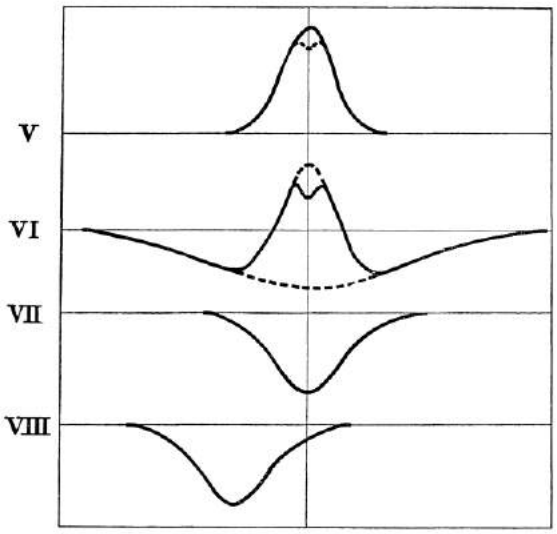

FIGURE 1.5: Secondary morphological classes proposed by Beals.

1.4 Thesis Outline

This thesis is structured as follows: Chapters 1 to 3 provide background on the problem statement, motivation, prior work and the data that were used in prototyping methods and deriving the results presented in the subsequent chapters. Chapters 4 to 6 provide detailed results as well as a comparison of this work to contemporary methods in the literature. The thesis concludes with Chapter 7, which places the results in context, provides a summarised commentary of the results and points out future directions of work. Conclusions have also been provided at the end of each chapter where relevant and appropriate.

A detailed account of manual and machine learning-based methods to date is provided in Chapter 2, highlighting the relative advantages and disadvantages of these methods. Chapter 3 provides background on the DR3 data set, data preparation, re-sampling and feature engineering. Chapter 3 will also briefly introduce the concept of casting spectra as ”time-series”,

which will be followed with a more detailed discussion in Chapter 4. Chapter 4 presents a novel data-driven approach that exploits this concept to identify, cluster and classify emission-line spectra. This chapter will demonstrate the efficacy of this approach on a sample of data from a recent study by Čotar et al. (2021).

Chapter 5 compares the approach developed in Chapter 4 with a recent and popular machine-learning method known as t-SNE. t-SNE was introduced as a suitable method to classify a broad range of spectral types in the GALAH survey. One of the conclusions of this chapter is that t-SNE is not sufficiently effective at identifying and classifying emission-line spectra. In particular, this work will demonstrate that the novel approach presented in Chapter 4 is more effective than t-SNE for the classification of P Cygni, inverse P Cygni and other emission-line spectra.

Chapter 6 builds on insights from chapters 3, 4 and 5 and introduce an end-to-end data driven machine learning method which ingests GALAH DR3 survey data, extracts emission-line spectra and classifies P Cygni, inverse P Cygni and other emission-line spectra. Using this novel approach, this work found over 7,000 emission-line spectra, over 200 P Cygni spectra and 200 inverse P Cygni spectra. The key results and ideas for building upon this work in the future are presented in Chapter 7.

2

A Review of Methods To Date

The GALAH survey selection function is defined by simple magnitude and Galactic latitude limits (Martell et al. 2017), and, as a result, it is expected that a majority of spectra are typical, reflecting the underlying Galactic stellar population. This places a constraint on the techniques that can be used to identify uncommon stars, such as those exhibiting emission-line spectra. Whatever methods employed, whether manual or automated, must be adapted to detect anomalies or outliers in data—for e.g., emission-line spectra—and separate them from more typical spectra. In this work, identifying emission-line spectra in large scale surveys is presented as a pre-processing step to classification and forms a critical step in the novel method presented in Chapter 6. This chapter presents essential background material and a review of methods that have been used in the past to identify Hα emission-line spectra, both for smaller-scale sets of observations during the latter half of the 20th century, as well as large-scale surveys in the early 21st century. These methods have also been placed in the context of their importance in the identification and classification of P Cygni, inverse P Cygni and other types of emission-line stellar spectra.

2.1 A Historical Perspective

Chapter 1 introduced the work of Beals (1953), who pioneered the study of emission-line stars in the 20th century. The work was manual and time consuming, and took the researcher literally decades to compile, with the assistance of a team that included secretaries and draftspeople. In spite of these efforts, the catalogue of data compiled in 1953 was small by modern standards. An in-depth literature search did not reveal further studies focused on identifying and classifying emission-line stars until several decades later, possibly due to the significant labour required to carry out these projects in the absence of modern data mining methods.

In 1993, an atlas of high-resolution line profiles with Hα emission-line spectra was published by Van Winckel et al. (1993). The authors noted that the radial velocities of the 59 emission lines considered were generally red-shifted compared to the underlying stellar velocities. They provide a classification scheme for Hα spectra based on emission-line morphology, devised entirely by using manual methods. More specifically, direct visual inspection and measurements of the width and shape of the line profile were used to classify these spectra, with wind velocity dispersion relative to the photosphere of the star used to determine the membership of spectra in each class.

A few of the classes identified by the authors include sub-type S1 (Figure 2.1), which displays narrow emission lines with no prominent absorption; sub-type S2 (Figure 2.2), which shows a clear absorption feature superimposed on a broad emission feature; and sub-type S3, which shows a strong absorption feature reaching at least the continuum level. Stars in sub-type S3 also have the smallest wind velocity dispersions, illustrating the relationship between the morphology of the spectra and the physical processes that generate them.

Following on three years later, Reipurth et al. (1996) studied the Hα emission-line profiles of pre-main sequence stars using manual methods. In addition to identifying T Tauri stars and Ae“Be stars using high resolution data (R~50,000), the study focused on the morphological properties of the spectra of P Cygni stars, as well as the physical processes that generate them. The study noted the discovery of complex morphological profiles among the T Tauri, Ae“Be and P Cygni star classes, for which the authors proposed a two dimensional classification scheme based on the relative height of a secondary peak compared to the primary peak, and whether the absorption line is blue- or redshifted. This has been reproduced in Figure 2.3. In their observed sample, the authors noted the classification of 25% symmetric profiles,

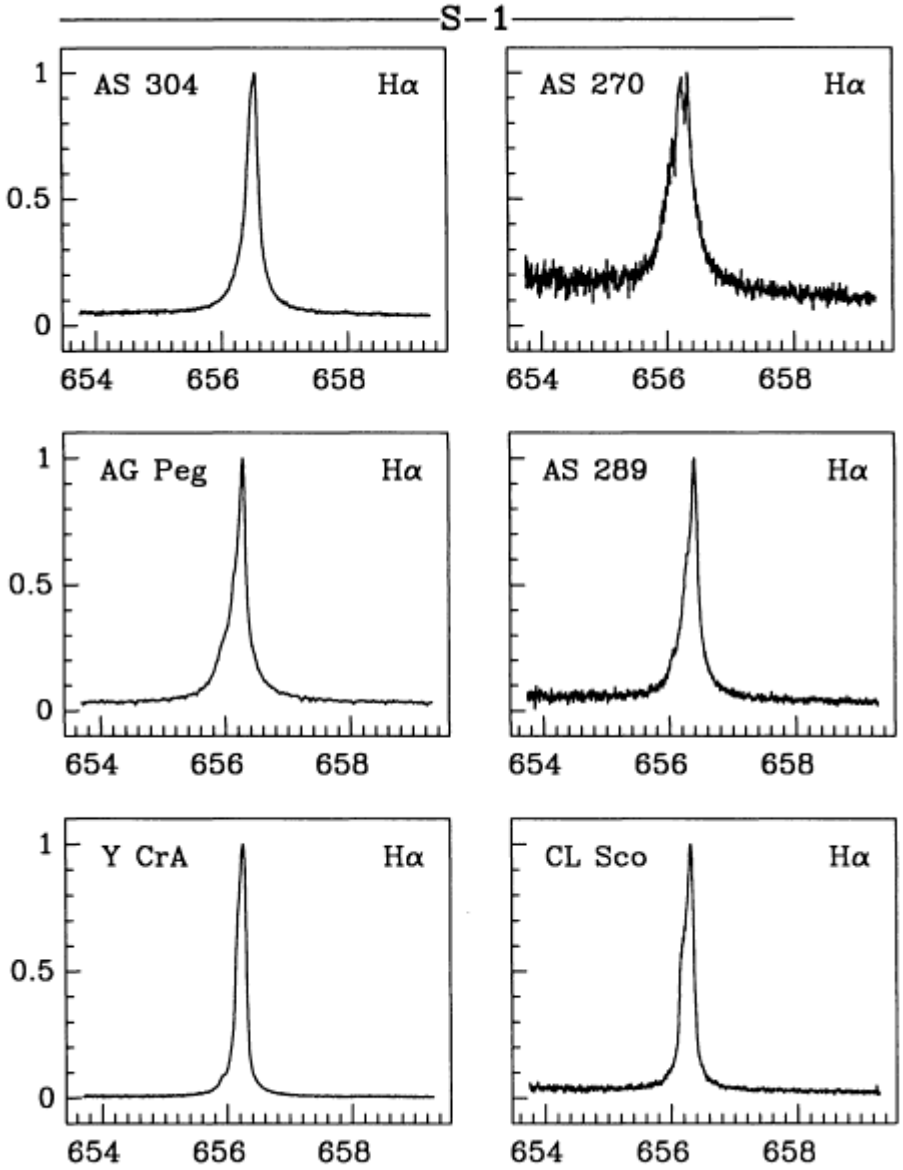

FIGURE 2.1: Sample spectra from sub-type S1 as classified by Van Winckel et al. (1993).

49% blueshifted absorption profiles and 5% P-Cygni profiles, with some 21% falling into the redshifted absorption category. In addition to this morphological classification, the authors also measured stellar wind velocities, with some stars recording extremely high velocities of ~900km/s.

The classification of P Cygni stars in Reipurth et al. (1996) follows the scheme proposed by Beals (1953). The authors also compared observational data to models in the literature, particularly models that constrain mass, radii and photospheric temperatures, although no

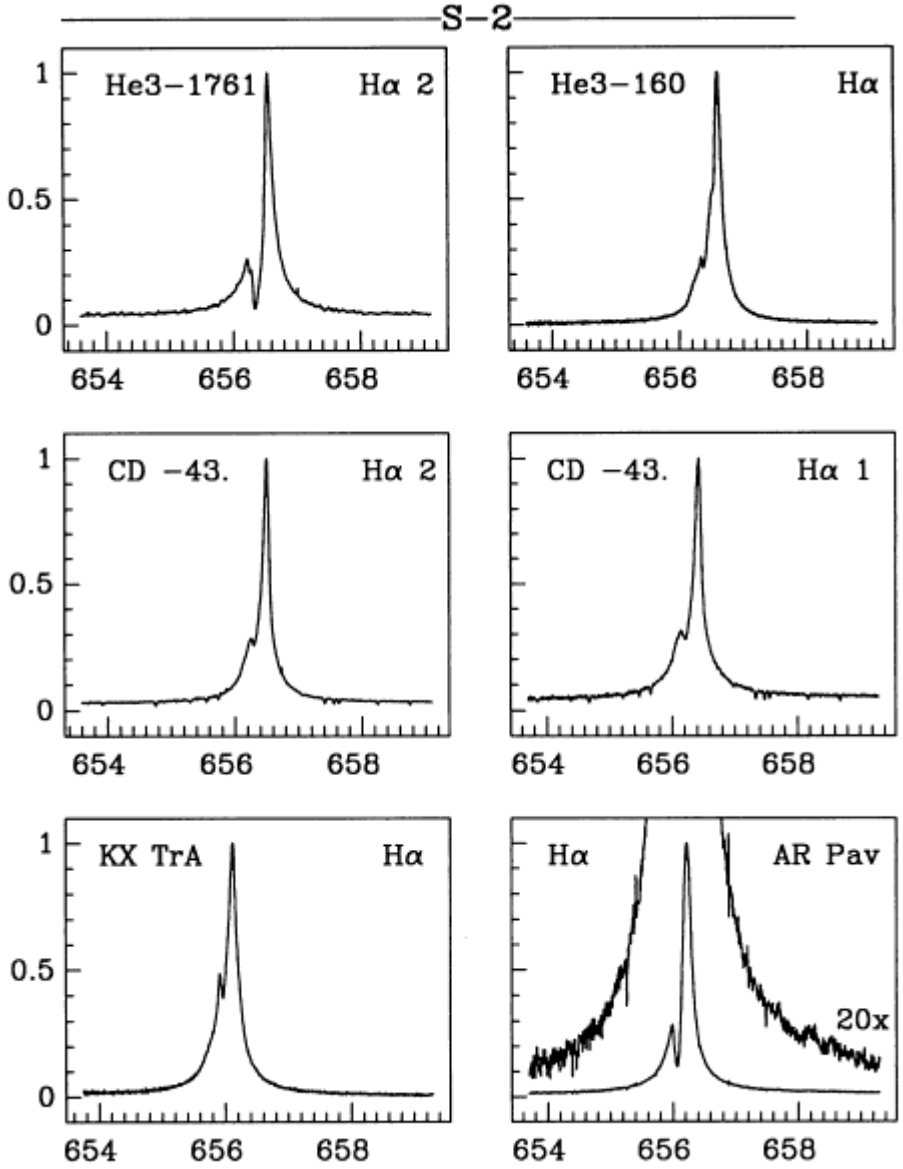

FIGURE 2.2: Sample spectra from sub-type S2 as classified by Van Winckel et al. (1993).

specific model details for P Cygni stars were provided. Further catalogues of Hα emission-line stars were produced by Kohoutek & Wehmeyer (1999). These catalogues contain 98 identified emission-line stars in the Northern Milky Way, but do not specifically identify P Cygni stars or inverse P Cygni stars.

Working with data from NGC 6611, Bonito et al. (2013) noted that for stars surrounded by active gas shells or optically thick winds, the morphology of emission lines could fall into categories such as symmetric with broad wings, asymmetric and, in extreme cases, P Cygni and inverse P Cygni. The authors used the classification scheme proposed by Reipurth et al. (1996) mentioned above, adhering to the type I - IV scheme with B and R suffixes to denote

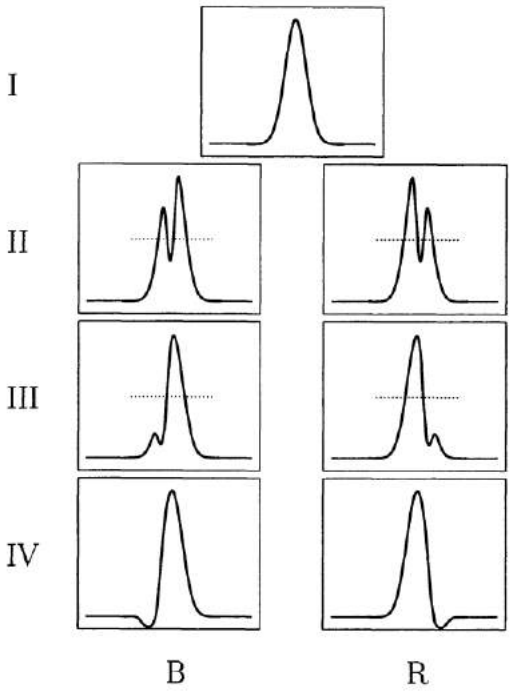

FIGURE 2.3: The morphology based classification scheme proposed by Reipurth et al. (1996) (reproduced). Depending on the location of the primary peak in relation to the secondary peak, the letters B and R—or blueshifted and redshifted, respectively—were appended to the Roman numerals.

blueshifted and redshifted emission lines respectively.

Traven et al. (2015) presented a catalogue of Hα emission-line stars identified in the Gaia-ESO survey. This survey's selection function was biased towards young open clusters, and consequently the authors noted a relatively large proportion of Hα emission-line spectra in their data: 3765 emission-line stars were identified from a sample of 22,035 spectra. This work is notable as it uses a combination of empirical rules and automated methods like spectral fitting to sort the Hα emission-line spectra into eight distinct morphological categories. These include single–component emission, emission blend, sharp emission peaks, double emission, P-Cygni, inverted P-Cygni, self–absorption, and emission in absorption. This work was briefly introduced in Chapter 1. A subset of these emission-line spectra are presented in Figures 2.4 and 2.5.

The Gaia-ESO survey conducted repeat observations of about half the identified Hα emission-line stars, and hence the authors were able to comment on the temporal variability of these stars. The conclusion was that while some morphological categories exhibited

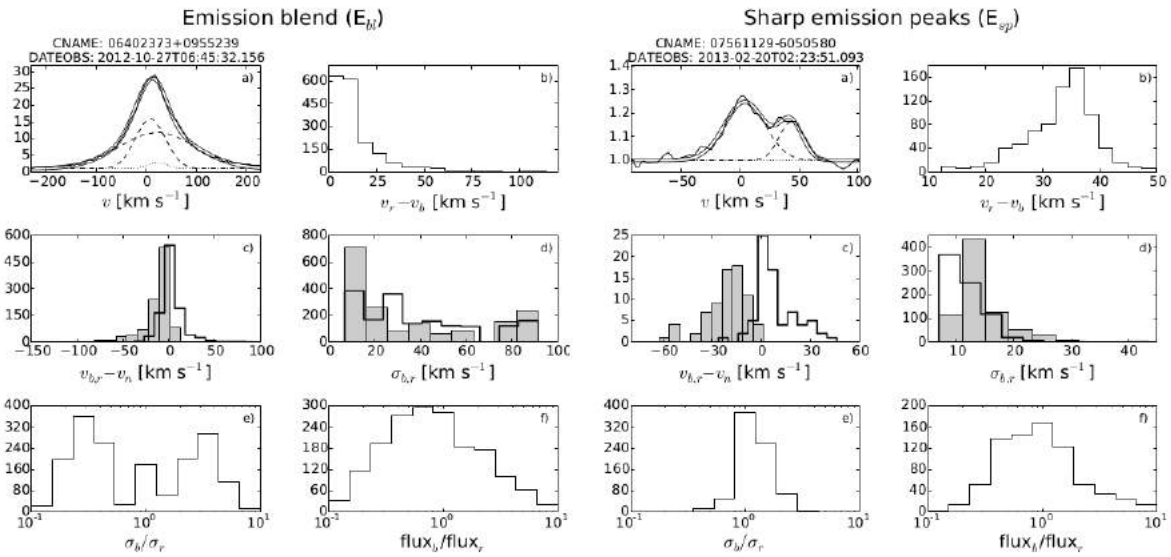

FIGURE 2.4: Classes of emission-line spectra identified in the Gaia-ESO Survey. Reproduced from Traven et al. (2015).

stability of their spectral profiles over time, P-Cygni and self-absorption profiles may not. Supplementary information on these spectra from cross-matches to SIMBAD, VizieR and ADS were also provided. In addition to these data, the authors provided wind velocity estimates based on a curve fitting procedure. In their discussion, the authors noted that the identification, classification and characterisation of emission-line stars can be valuable for automated pipelines in large surveys, where they can pinpoint outliers when calculating general stellar properties and abundances for the larger sample. Additionally, they note that the stars thus identified can be used for studies of star formation processes, interacting binaries and related fields of stellar physics.

Several conclusions can be drawn from this historical perspective:

1. These methods relied exclusively on visual inspection of spectra and manual methods to identify and/or classify Hα emission-line spectra. While this may have been a suitable approach in the past, it is extremely challenging to extend and scale these methods to data sets generated by million star all-sky surveys in the present day.

2. As noted in Chapter 1, a variety of spectral morphologies are believed to be generated by distinct physical phenomena linked to stellar winds and gas that surround stars. Thus morphology-based classification approaches, as demonstrated by Reipurth et al. (1996) and Beals (1953), are crucial to developing a greater understanding of stellar

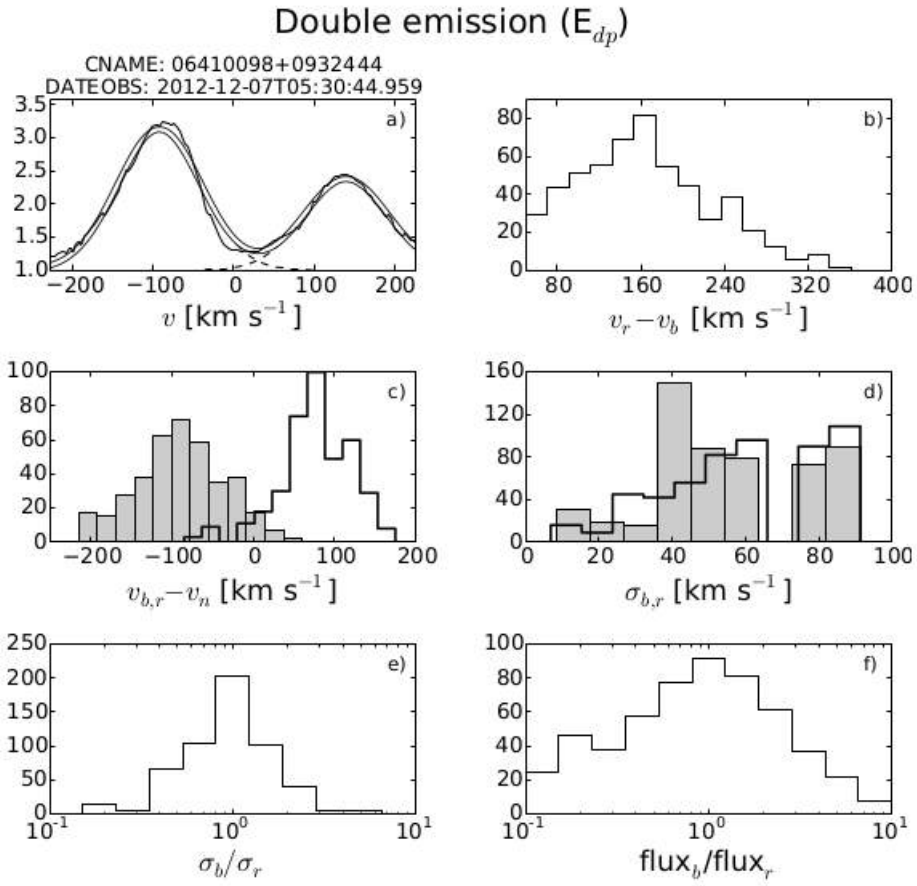

FIGURE 2.5: Classes of emission-line spectra identified in the Gaia-ESO Survey. Reproduced from Traven et al. (2015).

evolution and dynamics.

3. Finally, these studies identified P Cygni and inverse P Cygni (among other classes of spectra) as a subset of Hα emission-line spectra. The current work takes this fact to its logical conclusion, i.e., the probability (and efficiency) of identifying P Cygni and inverse P Cygni spectra can be increased if the search space and feature space of the raw data is reduced from a complete survey data set (e.g., the complete GALAH survey DR3 catalogue; Buder et al. 2021) to a much narrower subset of Hα emission-line spectra during pre-processing.

2.2 Recent Developments

The increase in data availability via large-scale spectroscopic surveys has necessitated the use of automated identification and classification methods. In recent years these methods have

included a veriety of statistical analysis and machine learning techniques. In general, machine learning approaches can fall into two categories: supervised and unsupervised learning. The former relies on the availability of a suitably robust set of training examples, while the latter attempts to generalise and learn from unlabelled data (Hastie et al. 2009).

While a full discussion and review of machine learning techniques is beyond the scope of this thesis, the methods that are relevant to this work are presented in subsequent chapters, with Chapter 4 in particular presenting a more detailed discussion on the relevant methods. The following section reviews four recent studies that use machine learning to identify spectra in large surveys, and discusses their strengths and limitations.

2.2.1 Applying K-means Clustering Directly on Data from APOGEE

The goal of k-means clustering is to partition a set of observations into a predefined number of clusters. Each observation would belong to a cluster with the nearest mean, which serves as the centroid or prototype of the cluster (MacQueen et al. 1967). Formally, given a set of n observations such as $x_1, x_2, ..., x_n$, where each observation is a d dimensional vector, the algorithm will partition the n observations into k sets $S = \{S_1, S_2, ..., S_k\}$, such that the intra-cluster variance is minimised. This objective can be represented as:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i \tag{2.1}$$

where μ_i is the mean of the points in S_i.

Using the identity,

$$|S_i| \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2 = \sum_{\mathbf{x} \neq \mathbf{y} \in S_i} \|\mathbf{x} - \mathbf{y}\|^2 \tag{2.2}$$

it can be shown that this is equivalent to minimising the pairwise squared deviations of points belonging to the same cluster.

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \frac{1}{|S_i|} \sum_{\mathbf{x}, \mathbf{y} \in S_i} \|\mathbf{x} - \mathbf{y}\|^2 \tag{2.3}$$

This method was used on high-resolution APOGEE infrared spectroscopic data (R~22,500), taken as part of the Sloan Digital Sky Survey (SDSS) (Blanton et al. 2017; Eisenstein et al. 2001). In the absence of labelled training samples in the APOGEE survey, Garcia-Dias et al.

(2018) used k-means to cluster similar spectra into distinct groups: each spectrum produced by APOGEE was treated as a d dimensional vector; the number of observations n was the number of spectra generated by APOGEE, which was approximately 150,000; and k was set to 50, presumably by a process of trial and error. The authors noted that they were able to separate dwarfs, sub-giants, red clump and red giant branch stars using this approach. The authors also commented that the approach is sensitive to initialisation, and thus sensitive to the number of clusters, k. One major limitation of this approach is that a discrete classification in flux space does not result in a neat organisation in parameter space, which implies that the authors were not able to link spectral features such as spectroscopic morphologies to the machine learning parameter space. The other limitation is the manual sorting of clusters that reduced the number of clusters from 50 to 9. This implies that certain spectra were incorrectly clustered by the k-means algorithm. Notably, the authors were unable to cluster emission-line spectra using this method.

The primary conclusion that can be drawn from this work is that k-means clustering, while robust on more traditional machine learning tasks, may perform poorly if it is used to cluster and ultimately classify morphologically-similar spectra such as P Cygni, inverse P Cygni and other emission-line classes traditionally identified using manual methods. A method that relates flux space, and consequently the morphology of the spectrum to a parameter space may perform better than k-means.

2.2.2 Combining Machine Learning and Manual Methods on Data from the LAMOST Survey

The LAMOST survey is a low-resolution spectroscopic survey with 10 million Milky Way stars as potential survey spectra. Zhang et al. (2021) were able to use a training and test set (labelled spectra) comprising of 5915 samples for spectral classification. This training set was based on data released by Hou et al. (2016), who developed the data set by using a combination of empirical rules and visual examination of 10,000 LAMOST spectra. The labelled data, including seven P Cygni and inverse P Cygni spectra identified by Hou et al. was used by Zhang et al. for supervised machine learning algorithms. Ten different supervised learning methods were then applied to this data set including including KNN (K-Nearest Neighbor), RF (Random Forest), AdaBoost, Naive Bayes (MultinomialNB, GaussianNB, BernoulliNB),

logistic regression, SVM (Support Vector Machine) and Artificial Neural Network (Single-hidden Layer, Three-hidden Layer). A comparison of the relative performance of these different methods was not provided by the authors, and a detailed discussion of all these methods would be beyond the scope of this thesis.

Zhang et al., however, note that the k-nearest neighbour and random forest methods outperformed all other methods. These two supervised machine learning models were then applied to 498,588 spectra, resulting in 56,574 potential Hα emission-line spectra. These spectra were then visually inspected, with a final candidate list of 30,048 Hα emission-line spectra. Despite the use of a number of machine learning methods, the authors fell back on manual visual inspection of spectra in building the training set, and during the classification of the identified potential Hα emission-line spectra. Thus it is clear that machine learning methods must be aligned with the desired end goals such that the work does not ultimately need to fall back on using manual methods.

2.2.3 Dimensionality Reduction in Action: Using t-SNE to Classify GALAH Spectra

Conducting computational operations such as clustering and classification on a higher dimensional vector space of size ~4500 can be challenging as it introduces significant computational overheads. In addition to the so-called curse of dimensionality, which will be detailed in Chapter 3, researchers would also face practical limitations due to the computational intractability of working on higher dimensional vector spaces. It would therefore be helpful to transform complex data sets from a high dimensional space to a low dimensional space while ensuring that meaningful features of the original data are preserved. In the case of P Cygni and inverse P Cygni spectra, these meaningful properties would presumably include some information about the morphology of the spectrum, although a reduction method may not guarantee this.

Principal component analysis (PCA) is arguably the most well-known dimensionality reduction method. However, PCA may not be suitable in the context of GALAH DR1, and certainly GALAH DR3, since these data sets are arguably biased towards non emission-line spectra. A PCA-led approach may select features that represent the absorption-line, while emission-line features are ignored. This hints at a possible outlier or anomaly detection-based

approach to emission-line spectrum identification. This idea is revisited in Chapter 6.

More recent and novel dimensionality reduction methods such as t-distributed stochastic neighbour embedding (t-SNE) (Van der Maaten & Hinton 2008) have been used on spectral data from GALAH DR1 (Traven et al. 2017). A stellar spectrum of vector length d (a wavelength grid of size d) can be used to create a vector space of dimensionality equal to d. In the case of GALAH DR1 this value is ~4500. Given a vector of size d, the application of the t-SNE algorithm will project this vector space onto a two dimensional vector space. The distances between data points on this two dimensional vector space can then be used to cluster similar data points into similar groups using a variety of popular clustering methods such as DBSCAN (Ester et al. 1996) or HDBSCAN (Campello et al. 2013). A detailed discussion of this method and its suitability to this work is presented in Chapter 5.

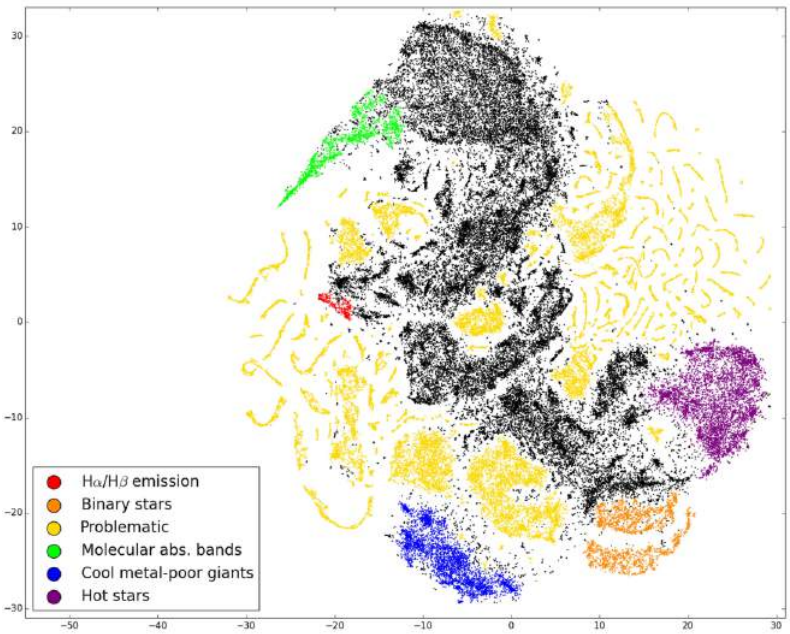

FIGURE 2.6: The t-SNE plot for GALAH DR1 with classified regions, reproduced from Traven et al. (2017). The x and y axes do not have a direct physical meaning but serve as spanning vectors for the two dimensional space.

Using this method, Traven et al. classified six distinct stellar types and identified Hα and Hβ emission-line spectra in GALAH DR1. These spectra were further examined and 18 P Cygni spectra were identified. The identification of P Cygni spectra was a sub-component of a broader study of stellar types in GALAH DR1. The authors noted that the number identified was lower than expected, which implies that there is significant scope to develop methods capable of detecting a larger number of emission-line stars in the GALAH survey. Crucially,

this method was not able to separate P Cygni spectra from double peaked spectra, emission superimposed on absorption, and other types of emission-line spectra. When examining the t-SNE plot (Figure 2.6), it is clear that the authors were not able to separate Hα emission-line stars as a distinct cluster from the unclassified region (black), but rather fell back on manual tagging of this region on the projection space.

Dimensionality reduction can overcome problems such as computational intractability but this should ideally not be at the cost of information losses related to the morphology of the spectrum. Since these morphologies uniquely identify the spectral types, a significant loss of this information will lead to poor classification performance. This point is revisited in detail in Chapter 5, providing examples using t-SNE where this type of information loss may have occurred.

2.2.4 Neural Networks as Anomaly Detectors: Using an Autoencoder on Data from GALAH and Other Surveys

In contrast to t-SNE, Čotar et al. (2021) introduced an anomaly detection-based approach to identify emission-line stars, the performance of which was a significant improvement over the t-SNE based method described above. This method used an autoencoder (AE), which is a type of artificial neural network (ANN) that takes input data and reduces it to a pre-selected number of ”latent features” that inhabit a low dimensional vector space, known as the latent space. The latent space can capture important features that exist in the original d dimensional vector space. By processing n samples of d dimensional vectors, the autoencoder can then learn the latent space representation of the higher dimensional features. This is known as encoding. In the next portion of the network, the network then attempts to recover the original data from the latent vector space. The process that reduces a d dimensional vector and vector space to a vector space of size $p ¡ d$ is a dimensionality-reduction method similar to PCA or t-SNE. The portion of the network that takes a p dimensional vector from the latent space and then projects it back to a d dimensional vector space is essentially the inverse process of the dimensionality reduction procedure. This component is known as a decoder. Since the original data are generated from the latent space, autoencoders fall into a class of ANNs called generative models (generative ANNs). Note that the latent space for an autoencoder is discrete. ANNs that can generate continuous latent spaces in this manner are known as

variational autoencoders—VAEs—but are beyond the scope of this work.

Autoencoders are widely used in anomaly detection applications (Sakurada & Yairi 2014) and can be suitably adapted to detect atypical spectra from survey data that contain a majority of typical spectra. If the autoencoder model is trained on non-anomalous data ("normal" data), the model will learn the latent space representation of non-anomalous data. Subsequently, if an anomalous data point is passed through the model, the autoencoder will attempt to generate these data from the latent space representation it has learned. Since the learned model was trained on non-anomalous data, the autoencoder will generate an inaccurate representation of the anomalous data. This leads to a prediction error, which can be used as a flag to detect time-series as well as non time-series anomalies in data.

As noted previously, the GALAH survey is a general all-sky survey that can be expected to generate a significantly higher proportion of non-anomalous spectra—specifically, since it is not biased towards young, violent, hot stars or stellar nurseries, we expect the vast majority of spectra to show "typical" (non-anomalous) stellar Hα line profiles, i.e., an absorption feature. In this regime, an Hα emission-line, and consequently, P Cygni and inverse P Cygni spectra are rare and anomalous. If an autoencoder is trained on "normal looking" spectra which do not show emission lines near Hα, it can be made sensitive to Hα emission-line spectra.

This method was used by Čotar et al. (2021) to detect Hα emission-line spectra in GALAH survey data. The authors chose a network architecture that reduces a d dimensional GALAH spectrum to a p dimensional latent space representation; here, $d = 4500$, while $p = 5$. Presumably, p was set to 5 to potentially capture the primary stellar parameters and represent them within the latent space. The intervening layers and the final architecture capable of detecting Hα emission-line spectra are presented in Figure 2.7. The authors did not further classify P Cygni, inverse P Cygni or other types of emission-line spectra using this method.

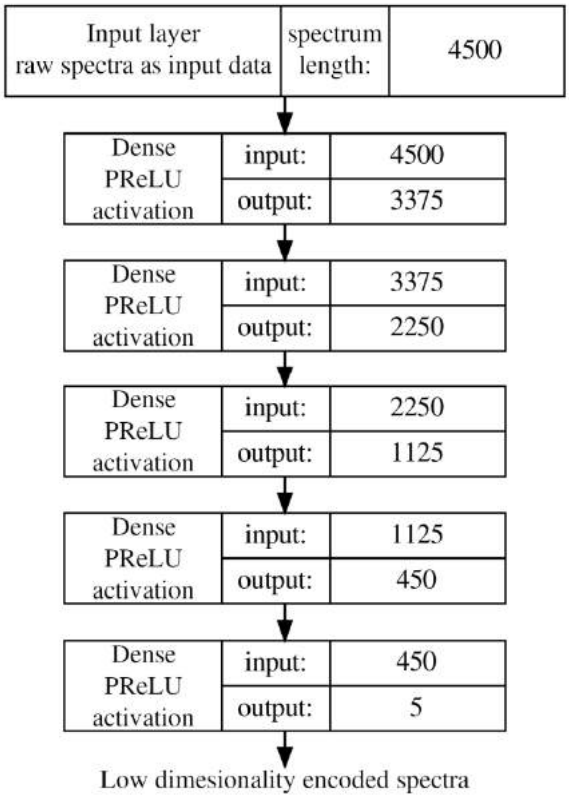

FIGURE 2.7: An autoencoder architecture capable of detecting Hα emission-line spectra in GALAH survey data. Reproduced from Čotar et al. (2021).

2.3 Selecting Emission-line Spectra Using Equivalent Width

It can be argued that emission-line spectra can be selected simply on the basis of computing the equivalent width of a spectrum within a region around Hα. In such a scheme, presumably, the emission-line spectra would betray themselves given that a strong emission feature may produce a negative value for the equivalent width. However this rather straightforward approach is not sufficiently robust for the purpose of this work for several reasons:

1. This method is not sufficiently sensitive to the multi-component nature of the emission-line spectra this work aims to classify. The equivalent width—essentially a single value—by itself cannot be used to identify P Cygni, inverse P Cygni and more complex spectral morphologies. In fact, for the aims of this study, methods that are sensitive to the morphological properties of the spectra must be prioritised over other methods. This is discussed further in Chapter 4.

2. This work aims to identify a wide selection of emission-line spectra, including weak emission-line spectra, which will not necessarily be identified by using the equivalent

width measurement method alone.

3. For weak emissions, the emission feature may only be visible as a reduced amplitude absorption feature (see Figure 1.5). However, the extent to which emission occurs in such a case cannot be accurately quantified using just the equivalent width. For this purpose, a reference spectrum or spectra may be required. This requires the creation of additional synthetic templates.

4. Finally, this method may not be sensitive to emission features that exist outside the equivalent width integration range, a real concern given the high wind velocities present in, e.g. P Cygni-type systems. However, this issue can be overcome by careful region selection.

Given these limitations, a simple equivalent width calculation would not be sufficient to discriminate between emission-line and non-emission-line spectra across the wide observational parameter space spanned by GALAH data. In subsequent chapters, this work will introduce a robust methodology to select emission-line spectra from the larger set of GALAH DR3, as well as review other methods used in the literature.

2.4 Concluding Remarks

It is clear from the review presented here that, given the scale of many contemporary astronomical data sets, modern work should rely heavily on automated methods when identifying atypical spectra such as those of emission-line stars. Attempts have been made at using machine learning methods to tackle this problem over the last five years with varying levels of success. If labelled data are available, supervised machine learning methods can be applied as in the case of Zhang et al. (2021). However, these methods must be applied appropriately if progress is to made with regards to relying on human beings to detect atypical spectra in large-scale surveys.

Regarding P Cygni stars and inverse P Cygni stars specifically, the literature dealing with their detection and characterisation is sparse and the available samples are few. A comprehensive catalogue of these spectra and stars does not exist at present. This can limit the use of supervised machine learning methods, and more importantly, can limit the science conducted on these spectra in the literature. Thus improvements to methods—and

the introduction of novel methods—can lead to more emission-line spectra being identified and classified, which in turn can lead to new science results on these stars.

The use of machine learning has been a relatively new development in the field. As far as can be determined from the literature, the use of t-SNE in 2017 is the first instance of applying machine learning to identify and classify Hα emission-line spectra. However, as will be demonstrated in Chapter 5, it can be challenging to use t-SNE to classify P Cygni and inverse P Cygni spectra, among other types of objects. Dimensionality reduction methods must be chosen carefully, so as not to lose too much information with regards to the features of the emission-line spectra.

Taking an anomaly detection or outlier detection approach such as Čotar et al. (2021), and using a neural network architecture such as an autoencoder can be beneficial in reducing the search space of surveys such as GALAH from the set of all spectra available to only the potential emission-line spectra. However this method cannot be used to classify particular types and sub-species of emission-line stars such as P Cygni and inverse P Cygni. Similar to dimensionality reduction, this method can reduce the complexity of the problem by allowing the researcher to focus on only the most relevant data points when attempting to identify and classify emission-line spectra. Combined with other methods, this method can serve as a pre-processing step when identifying and classifying emission-line spectra. This idea is explored more fully in Chapters 4 and 6.

Finally, popular unsupervised machine learning methods such as k-means, as well as supervised machine learning methods such as logistic regression may not be suitable. The latter requires labeled training data which was not available for this work, and there is evidence from the literature that the former performed poorly on the task of identifying and classifying emission-line stars in a large scale spectroscopic survey (Garcia-Dias et al. 2018). Hence these methods were not considered in this work.

3

The Data

This chapter presents an overview of the raw data used in this work, the challenges faced when working with higher dimensional data, strategies that were used to overcome these challenges and notes on data re-sampling and spectra region selection. It also presents a recent data set of emission-line stars that aided the development of the methods presented in subsequent chapters of this work.

3.1 Data Acquisition

This work utilises the most recent open access spectral data from the GALAH survey. At the time of writing, the GALAH survey is in its third data release (GALAH DR3; Buder et al. 2021). GALAH DR3 (hereafter DR3) comprises 678,423 spectra of 588,571 unique stars, of which approximately 80% are within a radius of 2 kpc. DR3 provides continuum-normalised spectra and errors for a majority of the observations (588,343 out of 588,571 total). In the case of problematic reductions, i.e., for spectra where this is not possible, object IDs and

flags have been provided. This study avoids using these problematic spectra.

The data are organised as individual `.fits` format files. Each file contains an object ID prefix (known as an `sobject_id`), followed by the last digit in the filename, which serves as a suffix for the camera number. GALAH uses the HERMES spectrograph at the Anglo Australian Telescope, which has four cameras, blue, green, red and infrared (Sheinis et al. 2014). Data from these cameras are denoted by the file suffixes 1, 2, 3 and 4, respectively. Thus, the file `1705090057010093.fits` is a data file for the object with `sobject_id=170509005701009` and contains spectral data from camera 3 (the red camera). An illustration of a reduced spectrum with data from all four cameras is provided in Figure 3.1. The HERMES red camera has a spectral range of 6478Å to 6737Å; this range is of particular interest for this work as it contains the characteristic Hα line at 6563Å. These individual files—totalling some 385 GB—were downloaded to a Macquarie University file server and were used as the data source for the work presented in this thesis.

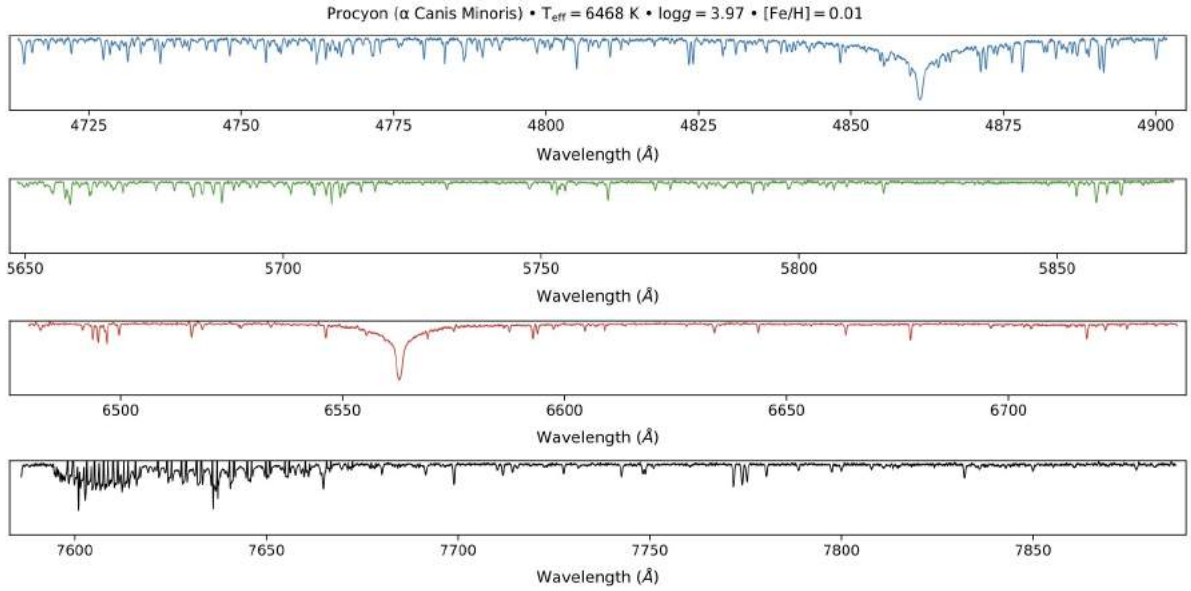

FIGURE 3.1: Normalised GALAH DR3 spectral data for all four HERMES cameras for the star α Canis Minoris.

This section provides an analysis of the feature space and search space of DR3 in the context of identifying emission-line stars. Feature exploration and engineering is a crucial step when conducting data analysis, particularly for the preparation of data prior to the application of machine-learning methods. Given that spectral features are recorded across four cameras at high resolution, the feature space of DR3 is significant.

As an illustrative example, consider only the red camera. The feature space calculation is as follows:

$$\lambda_{min} = 6478$$

$$\lambda_{max} = 6737$$

$$\Delta\lambda \approx 0.06$$

where $\Delta\lambda$ is the wavelength separation equivalent of the sampling rate of the wavelength grid. Thus the size of the wavelength grid is given by

$$N_\lambda = (\lambda_{max} - \lambda_{min})/\Delta\lambda \approx (6737 - 6478)/0.06 \approx 4317$$

This then gives the number of features from the red camera for a given spectrum,

$$N_f \approx 4317$$

Thus the total number of features for the red camera across all available DR3 data is

$$N_T \approx 4317 \times 678,423 \approx 2.9 \times 10^9$$

This calculation naively implies the existence of a billion-scale feature space, and consequently a potential billion-dimensional vector space. While a somewhat simplistic estimate, it nonetheless indicates that the data analysis and machine-learning strategy for tackling this data set should be planned and managed carefully. If care is not taken during feature engineering and pre-processing steps, the volume of data and complexity will lend itself to what is colloquially referred to as the ”curse of dimensionality”, which refers to the extraordinarily rapid growth in the difficulty of problems as the number of variables (or the dimension) increases (Kuo & Sloan 2005).

This work also utilises a recently published data set of Hα emission-line spectra. Čotar et al. (2021) used a prior version of the GALAH survey (De Silva et al. 2015) and data from two other surveys conducted with the HERMES spectrograph (the K2-HERMES survey (Wittenmyer et al. 2018) and the TESS-HERMES survey (Sharma et al. 2018)) to derive a catalogue of potential Hα emission-line spectra using a specific type of neural network known as an autoencoder. Combining data from the three surveys, this study used 669,845 continuum normalised stellar spectra as a data input source and included a small fraction of repeated observations.

sobject_id	ra (degrees)	dec (degrees)	Ha_EW (Å)
190206003501044	84.1585	9.68083611111	5.369496
190212001601342	83.8340416667	-0.703830555556	5.059437
161107003901377	82.9180416667	-66.7111666667	4.514134
190223001701241	82.5547083333	12.1460444444	4.3531947
190224002101351	85.4242083333	9.49156944444	4.1380215
190212001601351	84.2271666667	-0.418080555556	4.016535
190210001601238	76.7695416667	-2.59848333333	3.9660466
140610004401018	240.535208333	-22.9163611111	3.8685
160130003601074	87.5517083333	-17.9991666667	3.7972682
160422005701151	267.638125	-47.3779166667	3.7203794

TABLE 3.1: The 10 strongest Hα emitters identified by Čotar et al. (2021).

Čotar et al. identified 10,364 emission-line spectra with varying degrees of Hα emission components and sub-components. Summarised information on these spectra, including GALAH DR3 `sobject_ids` were released via CDS as open access data. These summarised data, excluding the actual continuum-normalised spectra, were presented as a single `.fits` format file. A subset of the ten strongest emission-line stars identified by Čotar et al. is presented in Table 3.1.

As detailed in Chapter 4, the Čotar et al. sample is used to demonstrate that a novel clustering approach can identify P Cygni and inverse P Cygni spectra and other species of emission-line stars. In addition, this data set is used to benchmark a well-established dimensionality reduction based clustering method called t-SNE, with these latter results presented in Chapter 5. Given the significantly smaller feature and search space, exploring and prototyping methods on this data set proved extremely beneficial when developing the full machine-learning method presented in Chapter 6.

3.2 Interpreting Spectra as Time Series

This work takes a unique view of continuum-normalised spectra from DR3, casting and interpreting these as one would a set of time-series data points. This proved extremely useful

as a mental model when developing the methods presented in this work. This section provides some insights into the motivation for using this conceptual model.

A plot of flux recorded by each camera, when presented as a function of wavelength, can be treated mathematically as traditional time-series data. While a monotonically-increasing time axis is not included in the DR3 data, the monotonically-increasing wavelength grid can serve as an analogue to the time axis. Morphologically, the variation of normalised flux against a wavelength grid is analogous to a variable plotted against a time grid.

The data analysis method developed in this work takes inspiration from this time-series approach. While perhaps unconventional, it is a powerful technique for analysing stellar spectra. Precedent for such an approach can be found in other scientific fields such as chemistry and nuclear magnetic resonance (NMR) spectroscopy, where NMR spectra are subjected to signal processing methods originally developed for time-series analysis (Nielsen 2019). Viewing spectral data in this manner allows for the exploration of signal processing methods that are typically reserved for, and employed on, time-series data. Notably, the use of dynamic time warping-based clustering was inspired by this conceptual model. These results are presented in detail in Chapter 4 and 6.

3.3 Data Re-sampling

The raw DR3 spectroscopic data do not have a uniform sampling rate. This means that the normalised spectra will not have a common wavelength grid and thus makes direct comparison of the morphologies of two spectra challenging. While this sampling rate is approximately $\Delta\lambda \sim 0.06$ Å in the red camera (Čotar et al. 2021), it varies around this value, particularly in the third decimal place (and varies even more from camera to camera). Furthermore, flux values may not be recorded for all spectra close to the wavelength limits of the red camera ($\lambda_{min} \sim 6478$ Å and $\lambda_{max} \sim 6737$) Å. Thus, when comparing spectra to each other, and their morphological features in particular, it was evident that all continuum-normalised spectra would have to be re-sampled to a common basis and thus a common wavelength grid. The advantage of this approach is that spectral features, particularly morphological features, can be compared against each other more effectively.

In order to isolate the red camera (camera 3) data, the following file operations were carried out. All filenames of the DR3 `.fits` format files were read into an array. Files

with the suffix ”3” were selected. Additionally the number ”3” was stripped from this sub array of file names to generate a list of `sobject_id` values. This significantly simplifies data querying and reading operations, as all standard query and file read operations rely on only `sobject_id` and not the combined file name that includes the `sobject_id` and camera suffix. The added advantage of this approach is that it automatically excludes `sobject_id` values for which red camera data do not exist. While a list of such `sobject_id` values is published on the GALAH survey website, this study did not require the use of this list as the procedure above infers these `sobject_id` values directly from the `.fits` filenames. This process results in a collection of 588,344 `sobject_id` values. This is lower than the 588,571 stars recorded by DR3, with the difference attributed to stars for which there are no normalised red camera data.

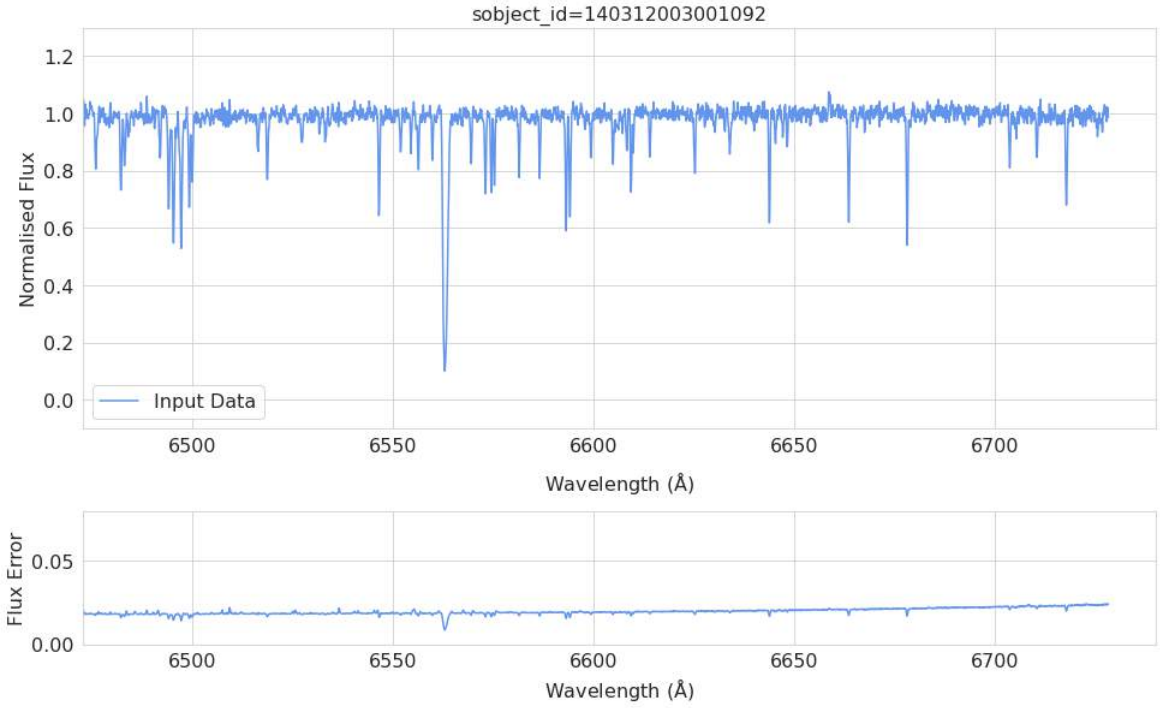

FIGURE 3.2: Red camera normalised flux data and error for `sobject_id` = `140312003001092` from GALAH DR3, prior to re-sampling.

To significantly minimise over-sampling and under-sampling, the spectra under consideration were interpolated to a common wavelength grid with a sampling rate equivalent to 0.06 Å. This rate is accurate to the sampling rate of each spectrum generated by the red camera to the second decimal place.

All red camera spectra were processed using the `spectres` Python package (Carnall

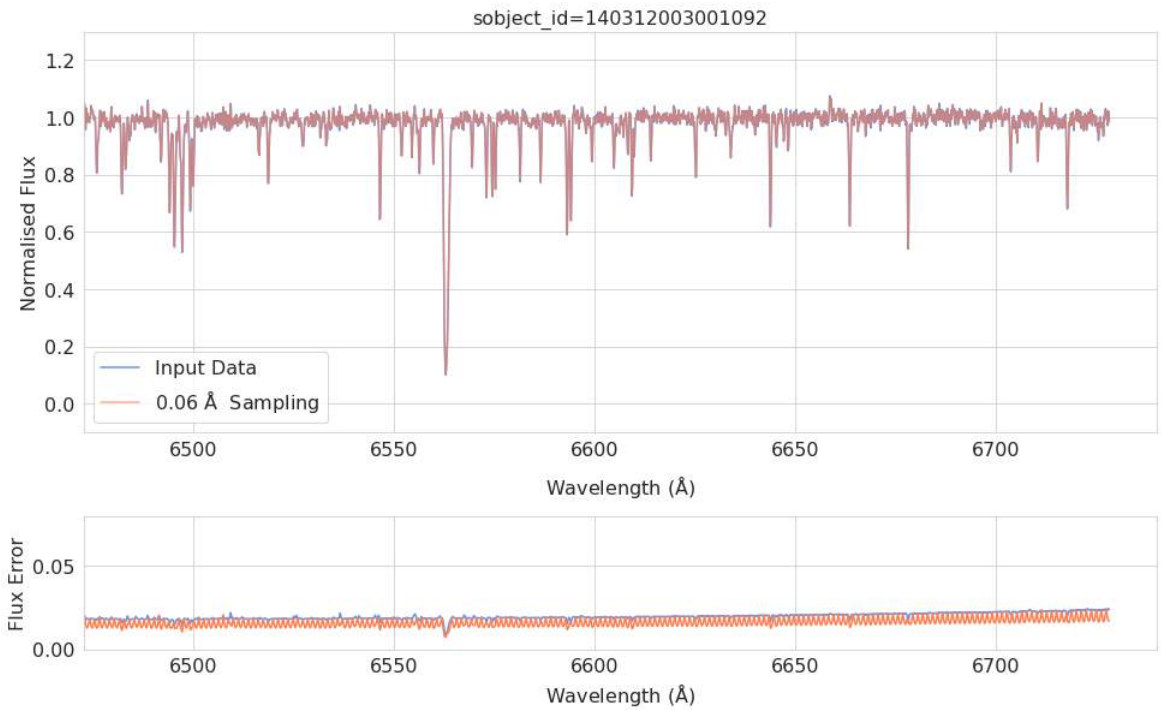

FIGURE 3.3: `sobject_id = 140312003001092`, re-sampled, using the method discussed in this section.

2017) to efficiently sample the following common wavelength grid:

$$\lambda_{min} = 6472.5$$

$$\lambda_{max} = 6740$$

$$\Delta\lambda = 0.06$$

This grid was chosen based on the range of wavelength separation observed in the raw data, as well as the work of Čotar et al. (2021). Normalised spectra and their errors from the red camera were processed in this way. The continuum value for normalised spectra in GALAH DR3 is set to 1. Thus, spectra for which flux values were not recorded at the tail and top end of the interpolated grid were padded with the value "1" to maintain the uniformity of the common wavelength grid. Re-sampling is computationally intensive and thus this process was offloaded to a university server. A representative example of this process is presented in Figures 3.2 and 3.3. For convenience, the resultant data and interpolated errors were saved as HDF5 files using the `.h5` file format in a single multi-dimensional array.

3.4 Region Selection

It was demonstrated previously that the total number of spectral features for the red camera $\sim 2.9 \times 10^9$. P Cygni, inverse P Cygni and other emission-line stars considered in this work show characteristic emission and absorption profiles near the Hα line. This fact can be used to significantly reduce the size of the feature space by selecting only the relevant region. The region of interest around Hα was selected to be 6561Å - 6565Å (Traven et al. 2017). By repeating the feature space calculation above, it can be noted that this region is sufficiently narrow enough to reduce the feature space by a hundredfold to $\sim 4.5 \times 10^7$, while simultaneously ensuring that it can encapsulate the emission features under consideration.

In code, this selection was implemented as a binary spectral mask which extracts the flux values for the relevant wavelength range while masking flux values outside this range. A masked version of the re-sampled data was stored in a separate `.h5` file for convenience. In terms of memory and disk allocation, this had the effect of reducing the memory footprint from ~60GB for the re-sampled red camera data to ~1GB for the Hα masked version of the same data. This improved the speed of read/write operations significantly.

3.5 Concluding Remarks

The following conclusions can be drawn from the analysis presented above:

1. When faced with higher dimensional data, dimensionality reduction and sample size pruning can be effective strategies in seeking to identify and classify atypical objects, such as emission-line star spectra.

2. Open data sets such as that provided by Čotar et al. can be useful as a prototyping aid to iteratively develop and evaluate machine learning methods, given the low volume of data compared to GALAH DR3.

3. Re-sampling can ensure that all spectra are compared on the same wavelength grid. This is particularly useful for developing machine learning methods, as the consistency of the data can be maintained throughout the process, particularly when comparing morphologies of spectra.

4. Region selection can significantly reduce the feature space and search space by efficiently limiting analysis only to those regions that are of interest for a particular study.

The following chapters will build on these conclusions, with Chapter 4 presenting a prototype unsupervised machine learning method that will identify and classify P Cygni, inverse P Cygni and other emission-line spectra in the data provided by Čotar et al.. Chapter 5 will compare the unsupervised machine learning prototype with t-SNE, while Chapter 6 will apply this method to the entire GALAH DR3 data set.

4

Developing a Framework for Classification

The volume and complexity of large spectroscopic data sets pose a significant challenge for developing methods to identify and classify emission-line stars. As previously explored, these and other constraints indicated that it would not be efficient or an effective use of time to trial multiple machine learning methods in pursuit of the end goals. As a result, a general framework and set of principles was developed based on the background provided in Chapters 1 to 3. These principles guided the development of a viable prototype method upon which the total machine learning method was built. The details of this framework, the methods that were selected, the *raison d'être* for these decisions and the results from the prototype are presented in the following sections.

4.1 Requirements and Constraints

As shown in prior chapters, morphological classification based on visual inspection of spectra cannot scale feasibly with the volume of data present in GALAH DR3. And while machine

learning methods such as t-SNE and autoencoders (Čotar et al. 2021; Traven et al. 2017) have been relatively successful in detecting Hα emission spectra, neither technique is able to further classify emission-line spectra into classes, such as P Cygni and inverse P Cygni.

When developing a framework for classification, it was found that an approach that is sensitive to the meaningful morphological differences between P Cygni, inverse P Cygni and other species is an important requirement. Given that the total is of size $\sim 2.9 \times 10^9$, a classification method must be able to overcome the "curse of dimensionality" and be computationally and memory efficient. As highlighted in Chapter 3, subsection 3.4, it is possible to overcome aspects of this by restricting the analysis to the region around Hα, and this will be expanded upon in subsequent chapters. Feature engineering must capture the understanding that P Cygni spectra exhibit a redshifted emission peak, while the inverse P Cygni spectra exhibit a blueshifted peak. A method should also be able to differentiate between classes such as double-peaked emission spectra and emission lines superimposed on absorption. This can be intuitive for a human being to do manually, but can prove challenging for a machine, as it must be trained on a well-constructed training data set that is sensitive to these features.

While emission-line spectra in GALAH have been identified in prior work, only a small number of P Cygni and inverse P Cygni spectra have been subclassified. For example, Čotar et al. were able to identify 215 emission-line stars in total from GALAH DR1, out of a sample of 300,000 spectra. The authors note that there are "relatively few" P Cygni spectra within the class of 215 emission-line spectra. Supervised classification methods require labelled training data per class. While the exact number of labelled samples required per class is highly problem dependent, it has been argued that a larger sample size ($\gtrsim 10^2$) is required for reasonable performance on applied machine learning problems (Halevy et al. 2009). Thus it is reasonable to conclude that, since DR3 does not have a sufficient number of labeled samples of P Cygni and inverse P Cygni spectra, a supervised learning approach to classification would not be suitable. When faced with labelled data scarcity, Halevy et al. for example argue that choosing an unsupervised learning approach can be significantly more powerful than using a supervised learning approach.

These conclusions narrowed the approach required to the unsupervised learning domain. Chapter 2 demonstrated that well-known unsupervised clustering methods such as k-means clustering generally fail to cluster and classify emission-line spectra (Garcia-Dias et al. 2018).

Given this result, this and related methods were not given further consideration. Instead, a first principles-based approach was adopted, with a focus on methods that are sensitive to morphological similarities and differences between spectra. More rudimentary methods, such as cross-correlation of signals, are examples of methods that could be sensitive to the various emission line morphologies. However, given that a labeled data set of the various classes of emission-line spectra in DR3 did not exist, this line of inquiry was abandoned as it was not possible to create mean spectra for convolving against all DR3 spectra for the various potential classes. Instead, methods were explored that created effects similar to pairwise cross correlation, provided a similarity score (or equivalent) with limited human intervention, and had minimum reliance on rule-based approaches that have been used in the past, for instance, Traven et al. (2015).

These constraints narrowed the search for suitable methods to a field known as unsupervised time series clustering, more specifically a clustering method based on a concept used extensively in signal processing called dynamic time warping (DTW) (Kruskal 1983). First introduced in 1975 as an algorithm for speech recognition (Itakura 1975), DTW has been modified and adapted extensively in various scientific disciplines such as signal processing, telecommunications and biomedical engineering. It was demonstrated in Chapter 2 that DR3 spectra, and indeed all spectra, are mathematically analogous to time series; while an individual spectrum does not contain a time axis, the monotonically-increasing wavelength grid serves as the analogue of the time axis. Time series methods such as DTW have been more generally developed in the domain of time series analysis (Nielsen 2019) and can be suitably adapted to stellar spectra. With clustering, this work did not rely on labelled data, but rather took a data-driven approach that learns from the emission-line morphologies present in the normalised spectral data present in DR3.

4.2 Dynamic Time Warping

Dynamic time warping (DTW) is a time series analysis algorithm that can measure similarity between two temporal sequences. The similarity can then be used to cluster morphologically similar spectra into meaningful groups. These clusters can then be used to classify spectra into distinct classes.

DTW is suitable for clustering problems where the morphology of the signal plays a

salient role over other features (Nielsen 2019) and where labeled data are unavailable. The name is inspired by the method itself, in which two signals are stretched or warped to align them on the temporal axis. For stellar spectra, this warping takes place in the wavelength domain and thus produces a discrete dynamic *wavelength* warping effect.

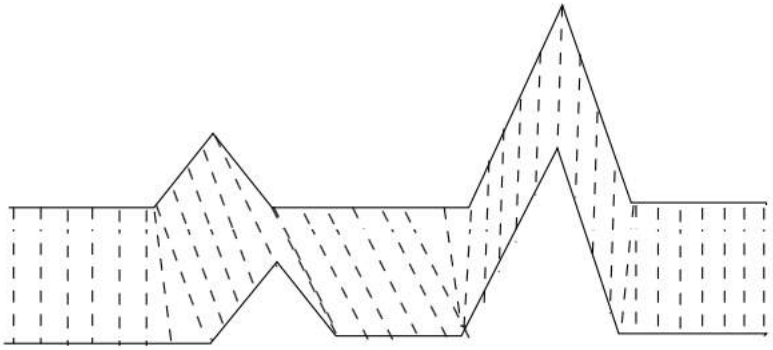

FIGURE 4.1: Each point on the wavelength grid is mapped to a point on the opposite spectrum, but there is no requirement that the mapping is one to one. Reproduced from Nielsen (2019)

As indicated in Figure 4.1, the algorithm works by expanding or contracting the wavelength axis to find the best alignment and thus ensuring that morphologically similar spectra can be compared. This algorithm is often described as being similar to comparing the shape of signals visually. The algorithm follows several steps and has several constraints which are as follows:

1. Every point on the spectrum must be matched with at least one point of the other spectrum.

2. The first and last indices of each spectrum must be matched with their counterparts in the other spectrum.

3. The mapping must be such that the wavelength is increasing rather than decreasing, i.e. the method should not match a point on one spectrum to a point on the other spectrum that has passed.

Steps 2 and 3 do not have a significant impact on the data set being used in this work as the data are sampled to the same wavelength grid and the spectra are thus of equal length.

There are many possible ways to align two spectra while adhering to these constraints. The algorithm chooses the alignment that minimises the distance between the spectra. This distance is a cost function and is measured as the sum of the absolute differences between matched points. The absolute difference in this context is the difference between the points' values (wavelength values). This distance measure then serves as the basis for clustering.

The Pythonic representation of this algorithm is as follows:

```
# Primary function
def distDTW(lambda1, lambda2):
    DTW={}
    for i in range(len(lambda1)):
        DTW[(i, -1)] = np.inf
    for i in range(len(lambda2)):
        DTW[(-1, i)] = np.inf
    DTW[(-1, -1)] = 0

# Calculate the optimum i.e. where distance is minimum
    for i in range(len(lambda1)):
        for j in range(len(lambda2)):
            dist = (lambda1[i] - lambda2[j])**2
            DTW[(i, j)] = dist + min(DTW[(i-1, j)],
                                     DTW[(i, j-1)],
                                     DTW[(i-1, j-1)])

 # Return the associated distance between two spectra
    return sqrt(DTW[len(lambda1)-1, len(lambda2)-1])
```

Despite the effectiveness of this algorithm, the computational complexity is still of order $O(N^2)$. As this presents a significant computational cost and overhead, the method developed here relies on a linear time complexity approximation—i.e., the $O(N)$ Python language implementation of DTW called `FastDTW`—to compute the optimum distance between spectra and thus the similarity (Salvador & Chan 2007).

In addition to computational complexity, an important consideration is available memory. Generally, all spectral data must be loaded into memory (RAM) when computing DTW distances. In the case of DR3, this implies holding $\sim 2.9 \times 10^9$ features in memory. The computational hardware available for this project had a memory capacity of $\sim$ 300 GB. `FastDTW` in particular required a memory capacity of at least 1TB for the $\sim 2.9 \times 10^9$

features. This is a significant amount of data to hold in memory and presented a serious obstacle.

Two strategies were developed to overcome these challenges and reduce computational and memory overheads. These are as follows:

1. Reduce the number of features. As explained in a Chapter 2, only the region around Hα is pre-selected in this work, and `FastDTW` was run only on this region.

2. Reduce the search space from the entire DR3 data set to a subset which has a higher chance of yielding P Cygni, inverse P Cygni and other emission-line spectra. The existence of Hα emission-line spectra is effectively a precursor to the existence of P Cygni and inverse P Cygni spectra. Thus the Hα emission-line data set from Čotar et al. (2021) was used for prototyping this method. This subset with $\sim$ 10,000 spectra includes data from GALAH and other related surveys with the HERMES spectrograph. It is assumed that this data set only contains Hα emission-line spectra, although Čotar et al. (2021) have indicated that this data set could contain non-emission-line ("typical") spectra as well, albeit as a comparatively minor proportion.

There are many machine learning methods that can be used to cluster groups of similar and dissimilar observations. These methods will generally compute method-specific distance measures as a similarity metric. DBSCAN, for example, computes a method-specific distance which cannot be overridden by a pre-computed distance metric such as a DTW distance (Traven et al. 2017). It is unclear whether a DBSCAN distance metric, or indeed a distance metric other than DTW, can capture the morphological features of emission-line stars. Due to time constraints, a thorough evaluation of the alternative distance metrics provided by the various clustering methods available was not undertaken. Instead, a clustering method which can accommodate pre-computed distances was chosen: agglomerative hierarchical clustering.

4.3 Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering is a method that can take a pre-computed distance metric such as a DTW distance and use it to cluster observations into classes. It is a well-understood and robust method that has proven to work well on time series problems relying

on DTW distances (Nielsen 2019).

With a similarity measure such as the pairwise DTW distances between spectra, hierarchical clustering can be used to group similar spectra into similar clusters. Once a similarity measure (or a dissimilarity measure) has been specified, hierarchical clustering produces a representation in which clusters at each level of the hierarchy are created by merging clusters at the next lower level. At the lowest level, each cluster contains a single observation. At the highest level, there exists a single cluster that contains all of the data (Hastie et al. 2009). There are two basic paradigms to traverse these levels (or tree), namely, aggolomerative (bottom-up) and divisive (top-down).

With agglomerative clustering, each spectrum will initially form a singleton cluster. At each step, the most similar spectra will be merged into a single cluster, producing one less cluster at the next higher level. The similarity between two spectra is based on the precomputed DTW distance between them. A lower distance implies a greater similarity, while a higher distance indicates a dissimilarity.

Based on the similarity (and dissimilarity) between individual spectra, a cluster dissimilarity can also be defined. Consider two clusters called A and B. The dissimilarity between the two clusters $d(A, B)$ is computed from the set of pairwise dissimilarities d_{ij}, where one member of the pair, i, is in A and the other, j, is in B. The complete linkage dissimilarity between the two clusters is set to be the dissimilarity of the furthest (most dissimilar) pair of spectra,

$$d(A, B) = \max_{\substack{i \in A \\ j \in B}} d_{ij} \tag{4.1}$$

This is also known as the farthest neighbour method. Other dissimilarity measures such as single linkage dissimilarity or nearest neighbour dissimilarity can be defined as

$$d(A, B) = \min_{\substack{i \in A \\ j \in B}} d_{ij} \tag{4.2}$$

Hierarchical clusters can be visualised using a dendrogram (Figure 4.2). A dendrogram is a binary tree that represents the recursive agglomeration (or division) of clusters. The height of the tree (or a branch) is proportional to the inter-group dissimilarity defined above.

This work used the complete linkage dissimilarity to cluster spectral groups. The justification for using this measure over the single linkage dissimilarity is to force the separation of P Cygni and inverse P Cygni spectra into distinct groups by exploiting the maximum distance between two individual spectra that belong to these groups. Other distance measures such

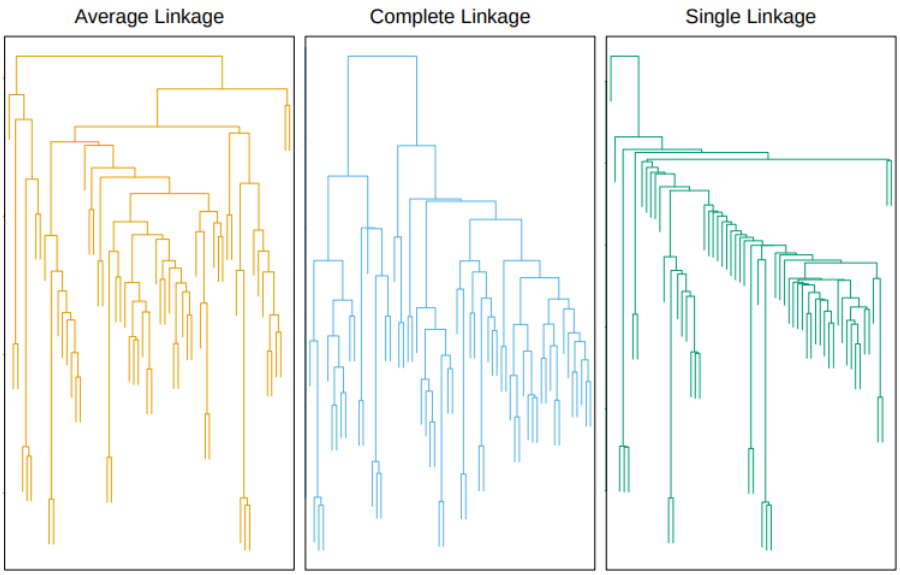

FIGURE 4.2: dendrograms for a given toy data set using different similarity measures. Note the longer branch lengths of the complete linked tree that selects for maximum dissimilarity. Reproduced from Hastie et al. (2009).

as group average clustering use the average dissimilarity between groups. Group average clustering can be less accurate in separating the data into clusters as it relies on an average spectrum per cluster, which may not completely capture the variation of spectral morphologies within a cluster. Therefore this measure was not considered, as it may be less accurate in separating P Cygni from inverse P Cygni and other emission-line spectra due to this averaging or smoothing effect. This work also relies on the agglomerative (bottom-up) method, as it is more robust and has been studied extensively in the literature compared to the divisive (top-down) method (Hastie et al. 2009).

4.3.1 Selecting the Number of Clusters

Given that the framework proposed above is an entirely unsupervised machine learning approach, it is generally not required that the number of clusters be specified in advance. In the absence of a predefined value for this parameter, the typical approach would require a plot of the dendrogram where suitable cuts can be made at a required level. However, since the tree can theoretically be cut at any level, and can have a maximum number of clusters equal to the number of samples and a minimum number of clusters equal to 1, a more meaningful and reliable cut can be made with the aide of astrophysical domain knowledge, as will be

demonstrated below. In the absence of such knowledge, this work would have to cut the dendrogram at each level, examine the clusters and subsequently decide on a suitable number of clusters. Furthermore, given that the sample size is at least 10,000 spectra, visually inspecting a dendrogram for this data set can be challenging. Where possible, this work used prior art such as Reipurth et al. (1996), Traven et al. (2017), and Zhang et al. (2021) to determine the number of clusters, thus eliminating the requirement to visually inspect a complex dendrogram of the order of thousands of branches.

Classes of Hα emission-line spectra have been found using manual methods and as such, prior work can provide some guidance regarding the number of clusters. Reipurth et al., in particular, proposed the existence of seven morphological groups which include P Cygni and inverse P Cygni, as well as five other Hα emission-line classes. However, if the number of clusters is set to a value of two, there is a significant risk that other morphologies will be included in the P Cygni cluster and inverse P Cygni cluster, thus leading to erroneously classified/labelled clusters. Thus, based on the prior art, this work took the number of clusters to be between six and ten as a suitable range. The number of samples were significantly higher than any data sets in prior art, especially the manual classification approaches detailed in the prior chapters. The justification to use a higher number of clusters such as ten is to account for the possibility of additional morphological classes that may have been missed in the prior art during manual classification. As it will be demonstrated in Chapter 6, this over-classification can be beneficial when working with more complex data sets.

4.4 Results

The results of the DTW distance calculation can be visualised as a distance cost plot. Figure 4.3 includes the pairwise distances for 6,977 samples from DR3 that are present in the Čotar et al. data set after the application of DTW. Note that the plot resembles a triangular matrix with the diagonal representing a value of zero for the self distance of a spectrum. Lower values indicate higher similarity while higher distance values indicate lower similarity (or greater dissimilarity). Zooming into a region as indicated in Figure 4.4 reveals further structure concerning similar and dissimilar spectra. The zero distance diagonal and the triangular nature of the distance matrix, are visible on this plot.

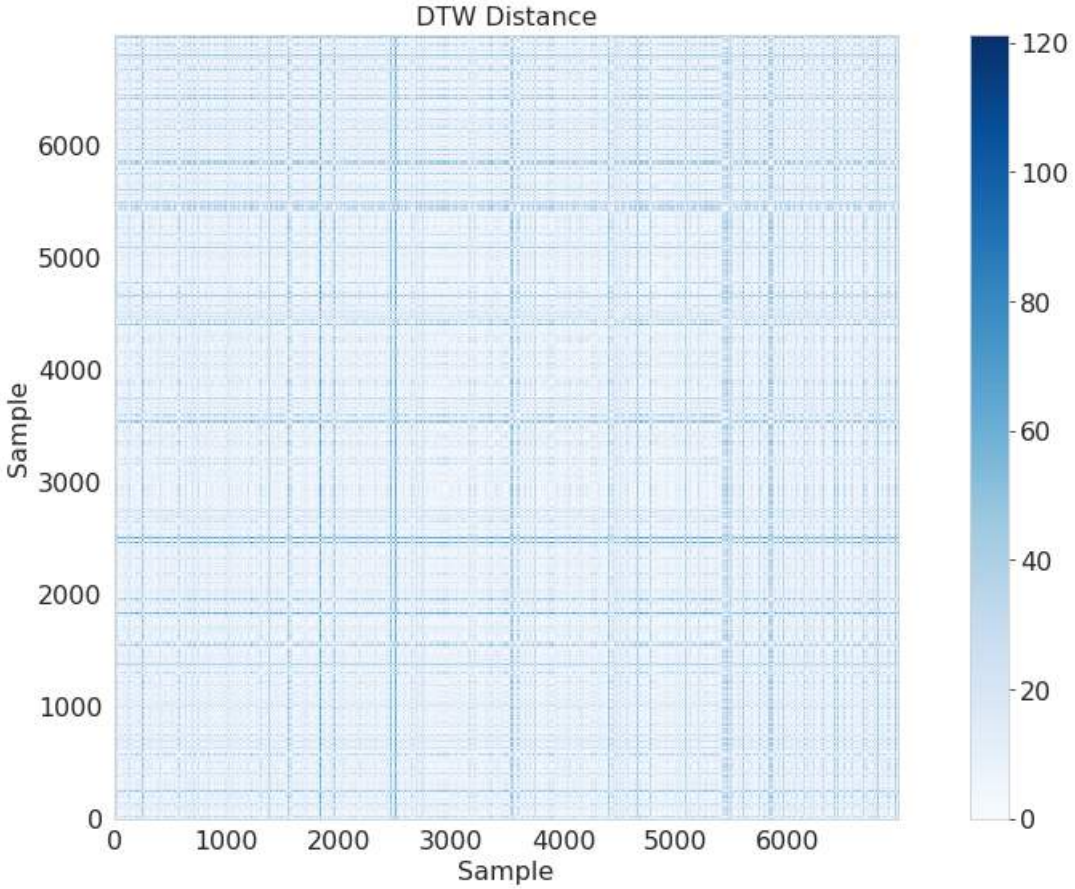

FIGURE 4.3: Pairwise DTW distances for spectral samples in Čotar et al..

This distance matrix was used as the basis for complete linkage aggolomerative hierarchical clustering. The number of clusters was varied between seven and ten. A mean silhouette score was calculated for each selection. This score was used as an additional selection criterion for the number of clusters. Silhouette scores can range between -1 and 1. Negative scores indicate that samples may be assigned to the wrong cluster. Values extremely close to zero indicate that clusters may overlap. The best achievable value is 1, although this is rare in practical unsupervised clustering problems. Thus the silhouette score is a measure of the efficacy of the clustering process.

In order to calculate the mean silhouette score, silhouette coefficients for all samples must be calculated as follows:

Compute the mean within-cluster distances given by a and then compute the mean nearest-cluster distance b for each sample. The sample silhouette coefficient is then given by

$$\frac{(b-a)}{\max(a,b)} \tag{4.3}$$

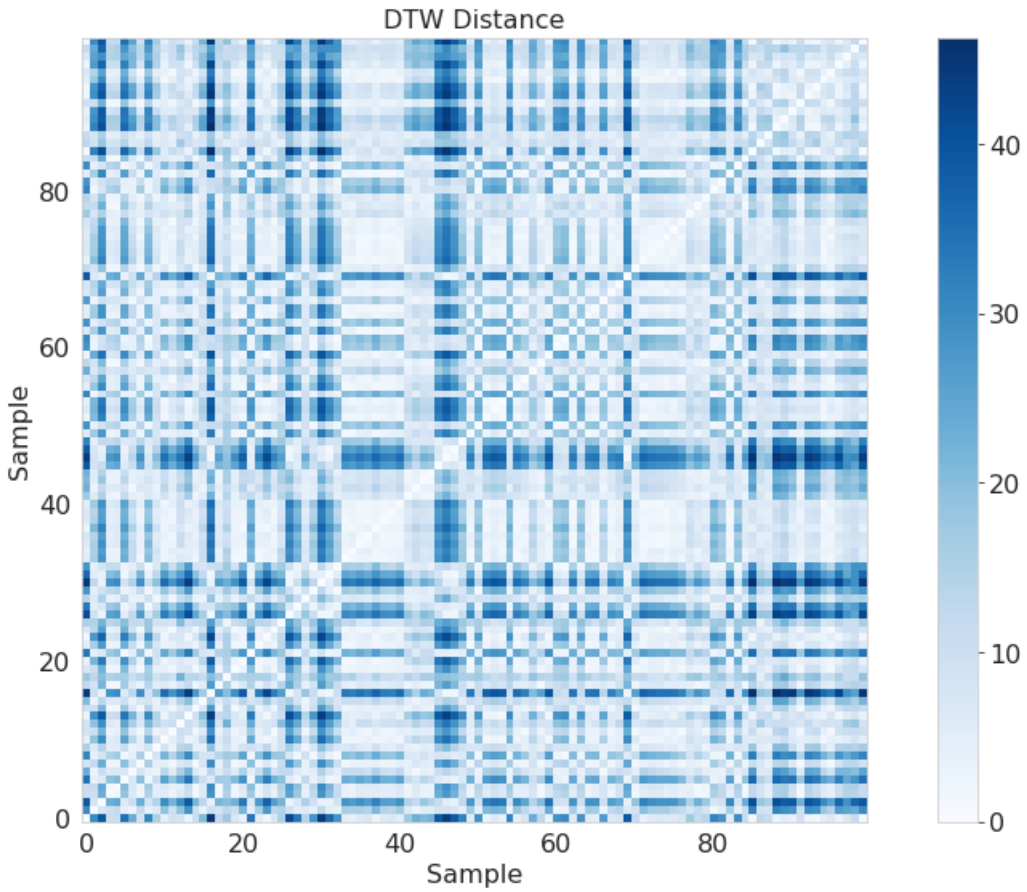

FIGURE 4.4: Pairwise DTW distances for spectral samples in Čotar et al. (zoomed).

Once the coefficient for each sample was calculated, the mean silhouette score for the entire sample was computed. Table 4.1 summarises the scores obtained.

At face value it appears that nine clusters is the optimum, given that this produces the largest silhouette score. However, upon closer examination it was concluded that ten was more suitable under the essential constraint that P Cygni and inverse P Cygni spectra should be adequately separated from other clusters. Thus, the spectra that belong to each cluster for both nine clusters and ten clusters were plotted. For nine clusters, while a P Cygni only cluster was identified, a cluster containing only inverse P Cygni spectra was not identified. However, when this parameter was set to ten, two clear clusters of P Cygni (Figure 4.5) and inverse P Cygni (Figure 4.6) spectra were identified.

Eight other clusters with various emission-line morphologies, such as double peak emission, were also identified. Other classes are presented in the Appendix. Given the silhouette score, it is possible that some P Cygni and inverse P Cygni spectra may have been misclassified (see Figure 4.6) and included in other classes. This can be addressed by further

Number of Clusters	Silhouette Score
6	0.2904
7	0.3033
8	0.3005
9	0.3092
10	0.3044

TABLE 4.1: Silhouette score comparison

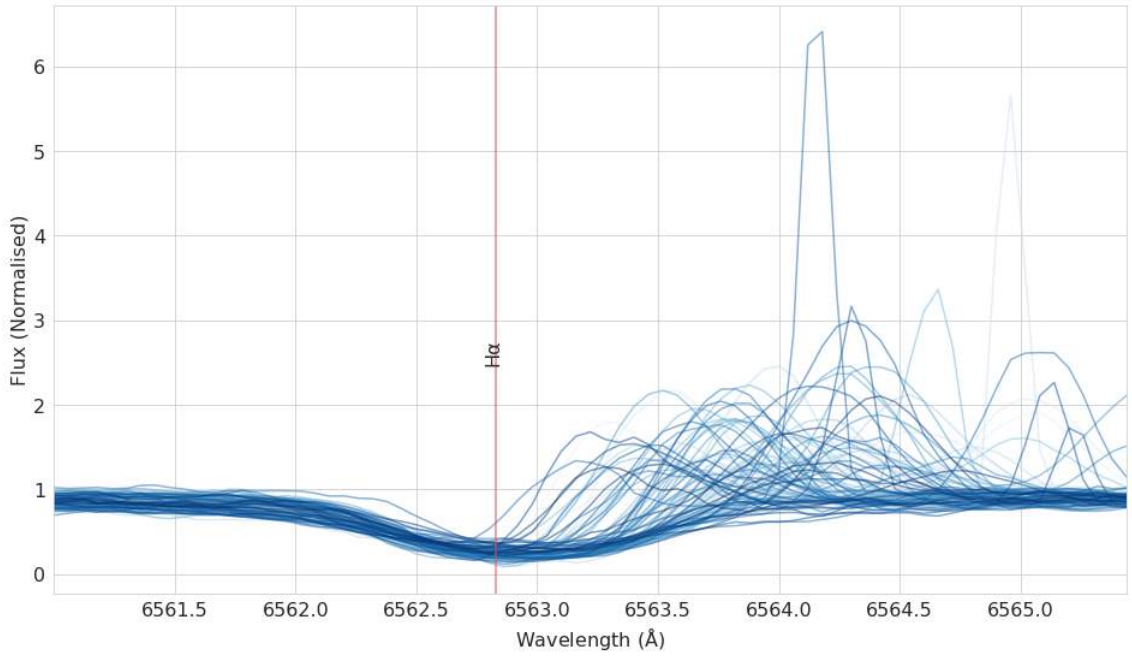

FIGURE 4.5: 102 P Cygni spectra identified using clustering.

sub-clustering and classifying spectra via a second pass of the scheme above. This work, however, did not progress with multiple passes of this method due to time constraints.

A sub-classification of the P Cygni and inverse P Cygni classes discovered in this process is presented in Chapter 5, where these results are compared to t-SNE. The advantages and disadvantages of this approach compared to t-SNE are also discussed in that chapter.

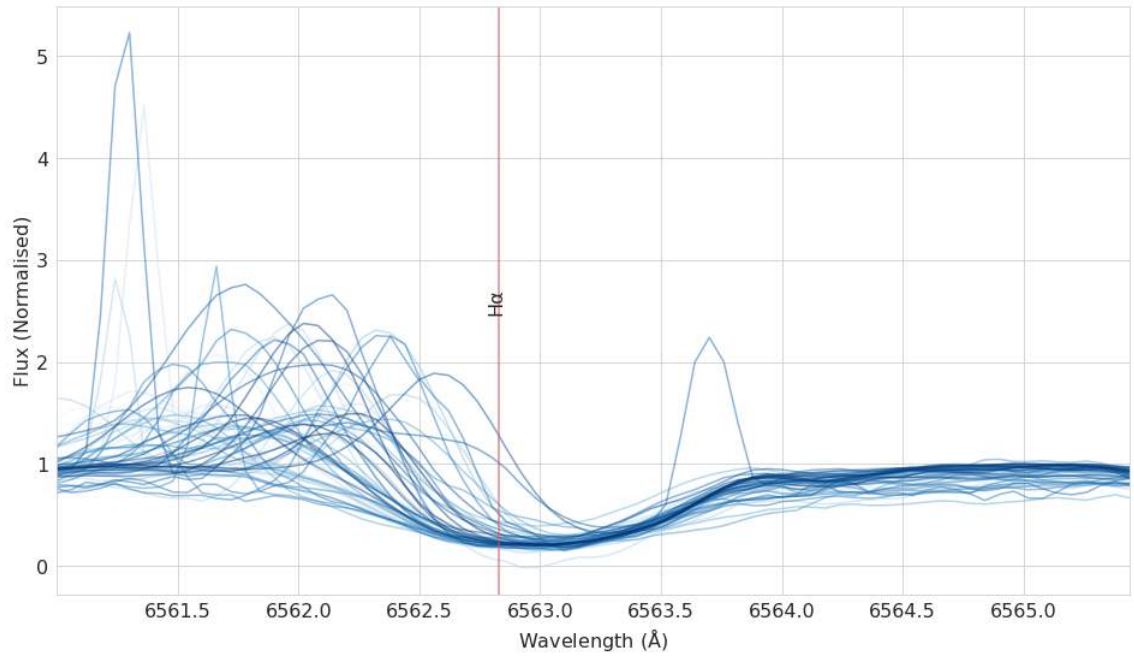

FIGURE 4.6: 62 Inverse P Cygni spectra identified using clustering.

4.5 Line Fitting

The P Cygni and inverse P Cygni spectra thus classified were modeled using a double Gaussian, with an offset for one of the Gaussian functions (Traven et al. 2015; Zhang et al. 2021). This mixture model with two Gaussians can be used to fit the line profile of P Cygni and inverse P Cygni spectra. Given that there are 102 P Cygni spectra that require fitting, this work adopted a semi-automated method where the initial conditions for the model parameters were driven by the data. Notably the minimum and maximum local flux values were used to set both the amplitude and mid-value of each peak and trough.

This work then utilised the popular Python based numerical optimisation package, `scipy` (Virtanen et al. 2020) to generate fitted models for the P Cygni and inverse P Cygni spectra using least squares optimisation. The results are presented in Figures 4.7 to 4.11.

The Gaussian mixture model is defined as,

$$f(x) = \frac{A}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right) + \frac{B}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right) + C \tag{4.4}$$

This seven parameter model contains an offset parameter C to account for the inverted Gaussian required to model the absorption trough/line of the P Cygni and inverse P Cygni spectra. The parameters A and B are to account for the respective amplitudes of the emission

and absorption lines. The uncertainties of this fit can be modelled more accurately within a Bayesian framework and the use of MCMC estimation (Hogg et al. 2010). A detailed discussion of this framework and estimation method is, however, beyond the scope of this work.

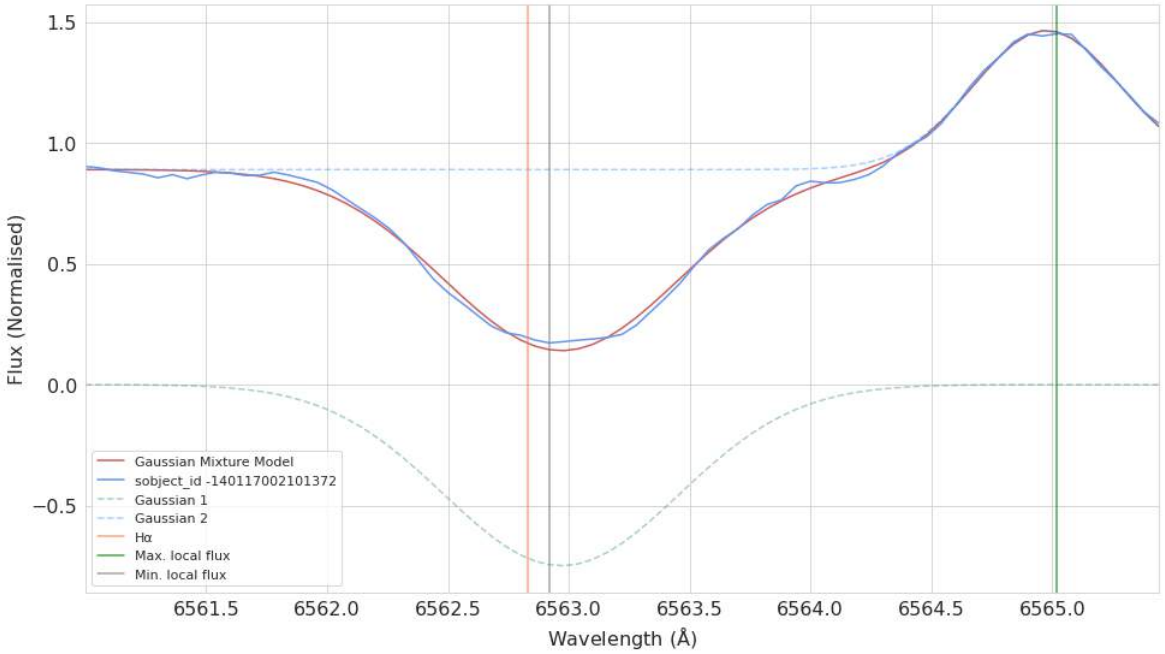

FIGURE 4.7: A Gaussian mixture model fit of one of the identified P Cygni spectra.

The line-fitting method described can be repeated for each P Cygni spectrum discovered by clustering. Presented below are a few examples of these fitted models. This technique can also extended to the inverse P Cygni spectra.

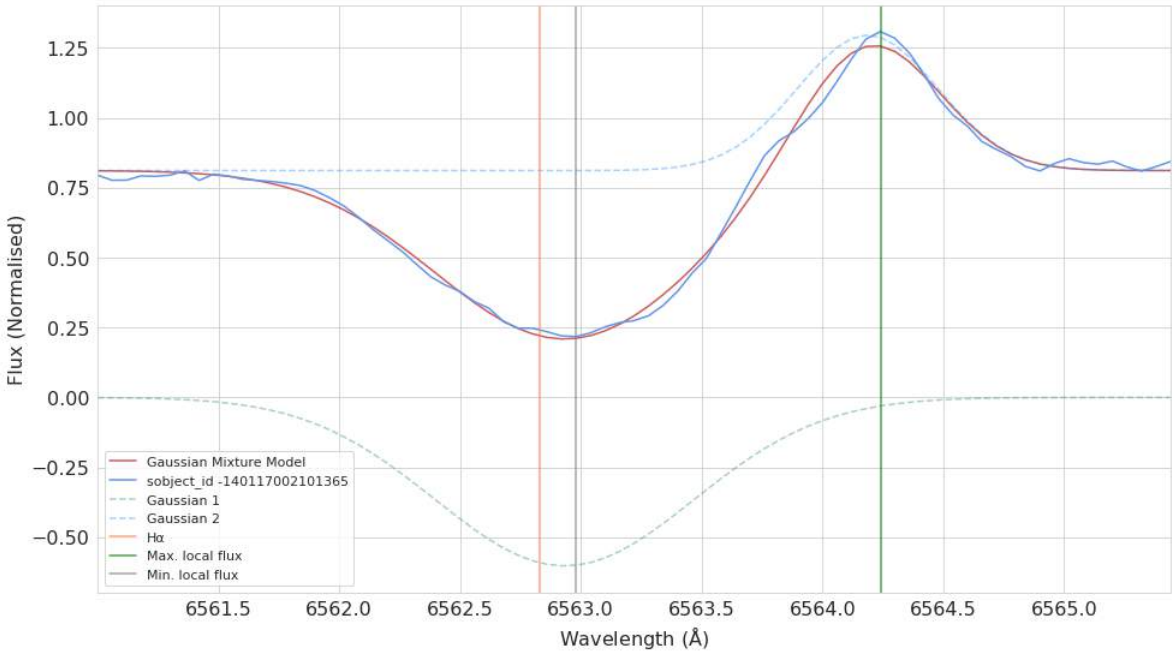

FIGURE 4.8: A Gaussian mixture model fit for sobject ID 140117002101365.

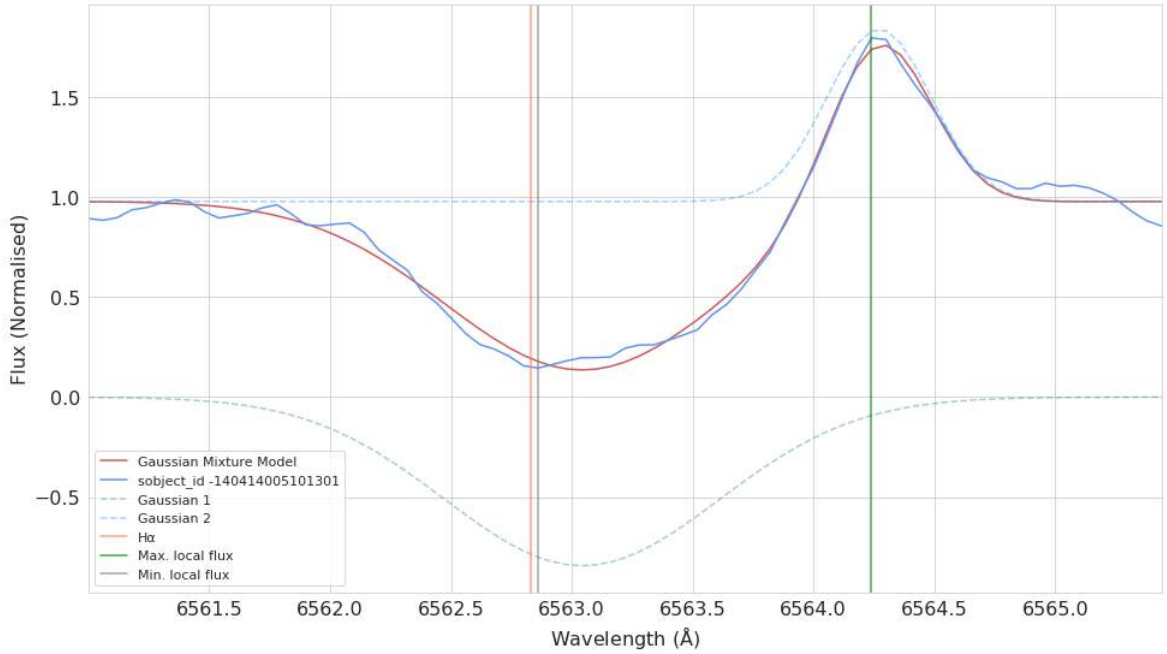

FIGURE 4.9: A Gaussian mixture model fit for sobject ID 140414005101301.

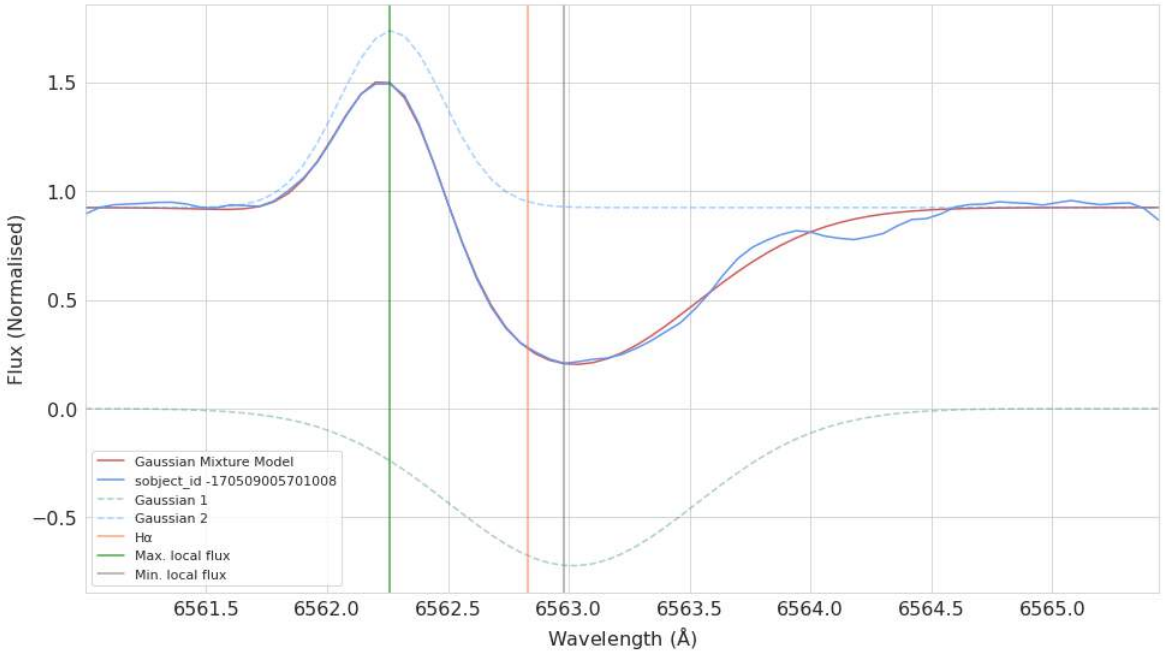

FIGURE 4.10: A Gaussian mixture model fit of one of the identified inverse P Cygni spectrum.

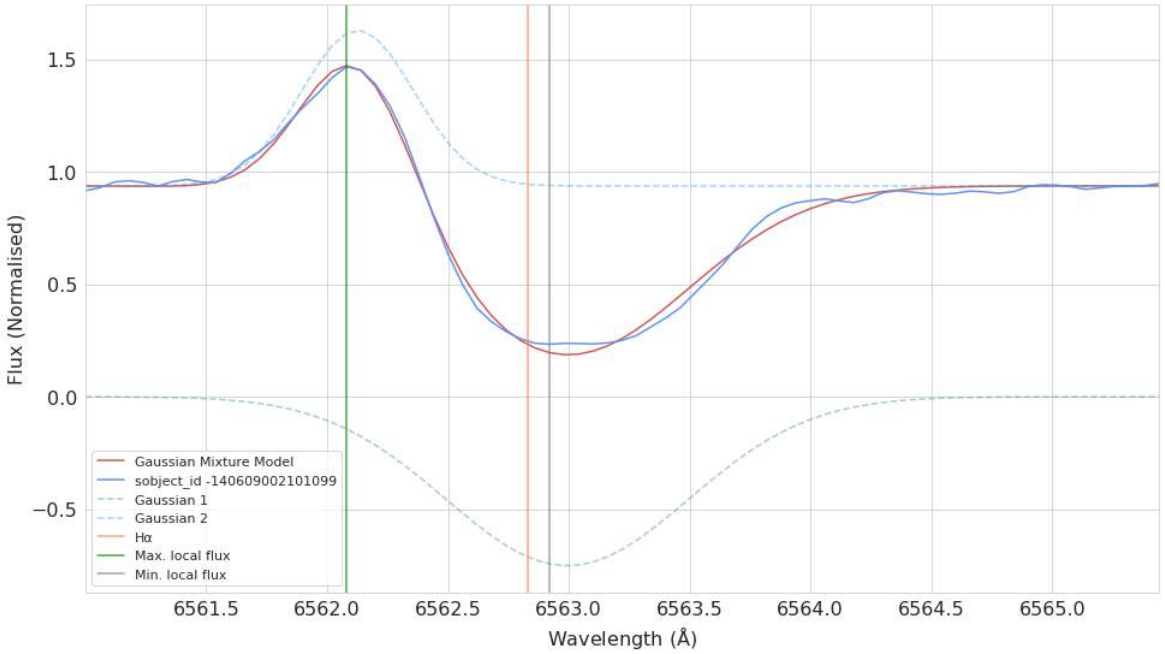

FIGURE 4.11: A Gaussian mixture model fit for one of the identified inverse P Cygni spectrum.

4.6 Concluding Remarks

Once a framework was developed, the sample data provided by Čotar et al. significantly accelerated the development of the methods discussed above. The primary reason for this is that methods can be tested and iterated at a more rapid pace given that the sample size was significantly smaller than GALAH DR3. Additionally, the sample data was well-pruned and presumably only consisted of emission-line spectra, although this sample purity may not be guaranteed. These conclusions make a strong case for the availability of open access data and code.

Dynamic time warping was able to learn from the data presented and provide pairwise distances that can capture the morphologies of the spectra. Combined with agglomerative hierarchical clustering, DTW can form the basis of a machine learning method that can identify, cluster and classify P Cygni, inverse P Cygni and other types of emission-line spectra.

Compared to methods such as k-means, logistic regression and t-SNE presented in Chapter 2, it can be argued that DTW is more sensitive to the morphological differences and similarities between spectra. In Chapter 2, t-SNE was also introduced as a method to identify Hα emission-line spectra in GALAH. Given the efficacy of DTW on a set of emission-line spectra provided by Čotar et al., this work considered whether t-SNE can be used as a pre-processing step in identifying emission-line stars prior to applying DTW. This hypothesis was tested, and the results are presented in the next chapter.

5

Evaluating t-SNE for Hα Emission-line Spectra Selection

The dimensionality reduction method t-SNE can map high dimensional data, such as high-resolution spectra, to a two dimensional feature space. Traven et al. (2017) used this method on GALAH DR1 survey data to classify spectra into classes such as binary stars, cool metal-poor giants and Hα emission-line stars. In Traven et al., manual intervention and visual inspection were used when identifying and classifying Hα emission-line stars.

Since Traven et al. used this method to identify a small number of Hα emission-line stars, it was sensible to investigate whether t-SNE could be used as a pre-processing step prior to applying the DTW-based method described in Chapter 4. In this chapter it will be demonstrated that, for GALAH DR3 spectra, there is no guarantee that this mapping or projection preserves morphological features. Thus it is not at all certain that this method is capable of identifying and classifying P Cygni, inverse P Cygni and other emission-line stars from a set of Hα emission-line stars, and hence the DTW-based method may be more

suitable.

5.1 Introduction to t-SNE

The dimensionality reduction method t-distributed stochastic neighbour embedding, or t-SNE, is a method that can be used to map higher dimensional data to a two dimensional plane (Van der Maaten & Hinton 2008). This t-SNE map can then be used to cluster and classify similar data points into groups and classes using clustering methods such as DBSCAN. Often, similar objects will be in close proximity to each other in the resulting lower dimensional space.

t-SNE was used successfully to separate and segment a subset of Hα emission-line spectra in the GALAH survey by Traven et al.; the authors isolated 215 Hα and Hβ emission-line spectra from a prior GALAH data release of approximately 300,000 spectra. A summary of these results are presented in Figure 5.1.

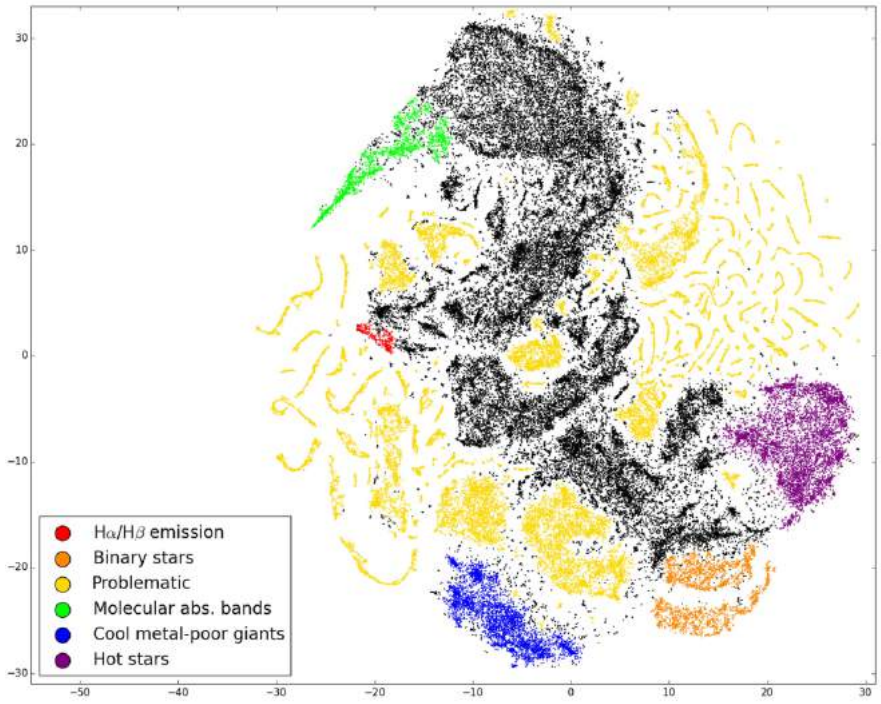

FIGURE 5.1: The t-SNE plot with classified regions reproduced from Traven et al.. The x and y axes do not have a physical meaning but serve as spanning vectors for the two dimensional space.

To overcome the so-called curse of dimensionality, high dimensional spectra from

GALAH DR3 can be mapped to a two dimensional plane. If it can be assumed that relevant features in the higher dimensional representation that are common to Hα emission emission-line spectra are preserved during the mapping from higher dimensions to two dimensions, then the spectra can be clustered into a distinct class. This class of Hα emission-line spectra can then be used as a starting point for further analysis using the DTW-based method discussed in Chapter 4. It is, however, not immediately clear whether this method is sensitive to the different morphologies of emission-line spectra. Thus the use of t-SNE must be evaluated if it is to be used as a method to select Hα emission-line spectra.

5.1.1 Dimensionality Reduction: The Mathematics of t-SNE

This section presents a brief mathematical introduction to t-SNE adapted from Traven et al. (2017). A more detailed discussion is found in Van der Maaten & Hinton (2008).

Consider N spectra where each spectrum is a higher dimensional object x_i. The low dimensional representation of this data is achieved by the optimal positioning of data points in the lower dimensional projection map. In order to achieve this, the t-SNE process defines a similarity between data points in the original higher dimensional space X and in the lower dimensional project space, Y (two dimensional for the purpose of this work). These are described by the symmetric joint-probability distributions P and Q respectively.

The pairwise similarity between data points x_i and x_j is modeled by the probability that one data point would pick another data point as its neighbour. This depends on the probability density under a Gaussian in space X while a Student's t-distribution is used in space Y. t-SNE first computes probabilities p_{ij} that are proportional to the similarity of spectra x_i and x_j as follows,

For $i \neq j$,

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)} \tag{5.1}$$

Note that $p_{i|i} = 0$ and $\sum_j p_{j|i} = 1$ for all i. Now define,

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{5.2}$$

and note that $p_{ij} = p_{ji}$, $p_{ii} = 0$ and $\sum_{ij} p_{ij} = 1$.

The bandwidth of the Gaussian distribution, σ_i, is set such that the perplexity of the

conditional distribution equals a predefined perplexity. This perplexity is a user defined quantity and is considered a hyper-parameter of this method. This implies that the bandwidth is adapted to the density of the data. Smaller values of σ_i are thus used in denser parts of the data space X.

Next, t-SNE will attempt to learn a lower dimensional map with objects y_i in Y that reflect the similarities p_{ij} computed above. The process measures similarities q_{ij} between two points y_i and y_j using a similar approach.

For $i \neq j$,

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}} \tag{5.3}$$

Here $q_{ii} = 0$ and the t-distribution with one degree of freedom is used to measure similarities between the lower dimensional points. Thus, dissimilar objects will be projected further apart on the map Y.

In order to determine the locations y_i on the map Y, the nonsymmetric Kullback–Leibler divergence shown below is minimized using gradient descent. The result of this optimization scheme is a lower dimensional map of the spectral data.

$$\mathrm{KL}\,(P \parallel Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{5.4}$$

This work used the t-SNE implementation from the popular Python package `scikit-learn` to generate all t-SNE mappings and results. While other implementations such as multi-core t-SNE exist (Ulyanov 2016), the performance gains of using processor optimised versions of t-SNE over the `scikit-learn` implementation are modest in this context. `scikit-learn` has a more substantive body of literature and is more efficient to use (Hackeling 2017). A full comparison of these methods are beyond the scope of this thesis.

5.2 Can t-SNE Be Used to Classify Emission-line Spectra?

In order to evaluate whether t-SNE is a suitable method that can identify emission-line spectra, this work used an identical approach to Traven et al. on data from DR3 to attempt to isolate, cluster and classify potential Hα emission-line spectra. Such emission-line spectra can then be subjected to the pipeline proposed in Chapter 4. The first step of this approach involves using t-SNE to dimensionally reduce spectra to a two dimensional t-SNE map. The spectra

were masked to select the Hα region only (Traven et al. 2017). The perplexity parameter was set to 30 as this figure provided reasonable separation of clusters in Traven et al.. Figure 5.2 contains the result of this t-SNE projection for DR3 data.

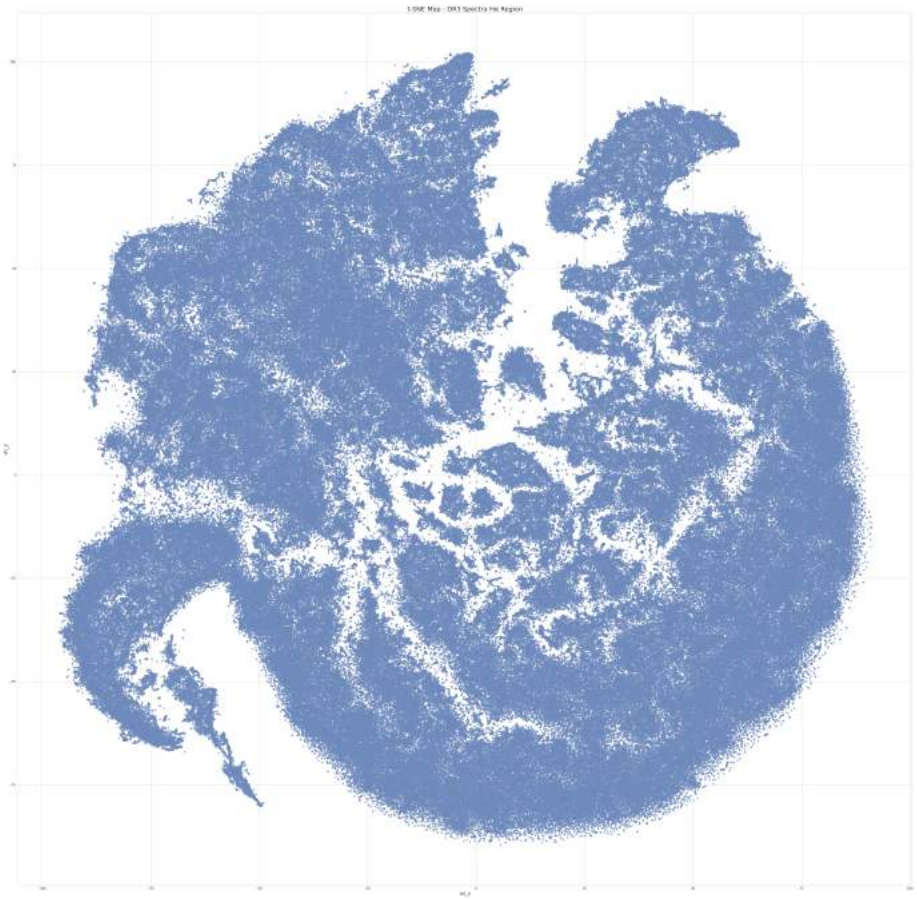

FIGURE 5.2: t-SNE map for all spectra in DR3. Each point is a two dimensional representation of a spectrum. Note that only the region around Hα of each spectrum was used to generate this map. As with the other t-SNE maps, note that the x and y axes do not represent physically meaningful quantities.

In order to assess whether the t-SNE plot can adequately separate Hα emission-line spectra, the validated Hα emission-line spectra discovered by Čotar et al. were marked on the same map by using the object IDs. These 7786 spectra are indicated in pink in Figure 5.3. While a portion of these emission-line spectra are clustered in the bottom left quadrant of the map, the other spectra remain scattered across the map with no obvious meaningful cluster that can be separated out of the t-SNE map.

Hence, following the distance-based clustering method used by Traven et al. it was not possible to cleanly separate the 7786 Hα-emission stars identified by Čotar et al.. This

reinforces the hypothesis that t-SNE may not be able to effectively capture morphological features from the higher dimensional space to the two dimensional plane. It may be possible to achieve better clustering performance by iteratively tuning the perplexity hyper-parameter and visually inspecting the map. However, due to time constraints, this was not attempted, as a suitable range for this parameter in the context of GALAH DR3 is not obvious. Future work is needed to study this method further, but again this is beyond the scope of this thesis. The question of whether t-SNE can perform *classification* of P Cygni and inverse P Cygni spectra better than the DTW-based method discussed in Chapter 4 is addressed next.

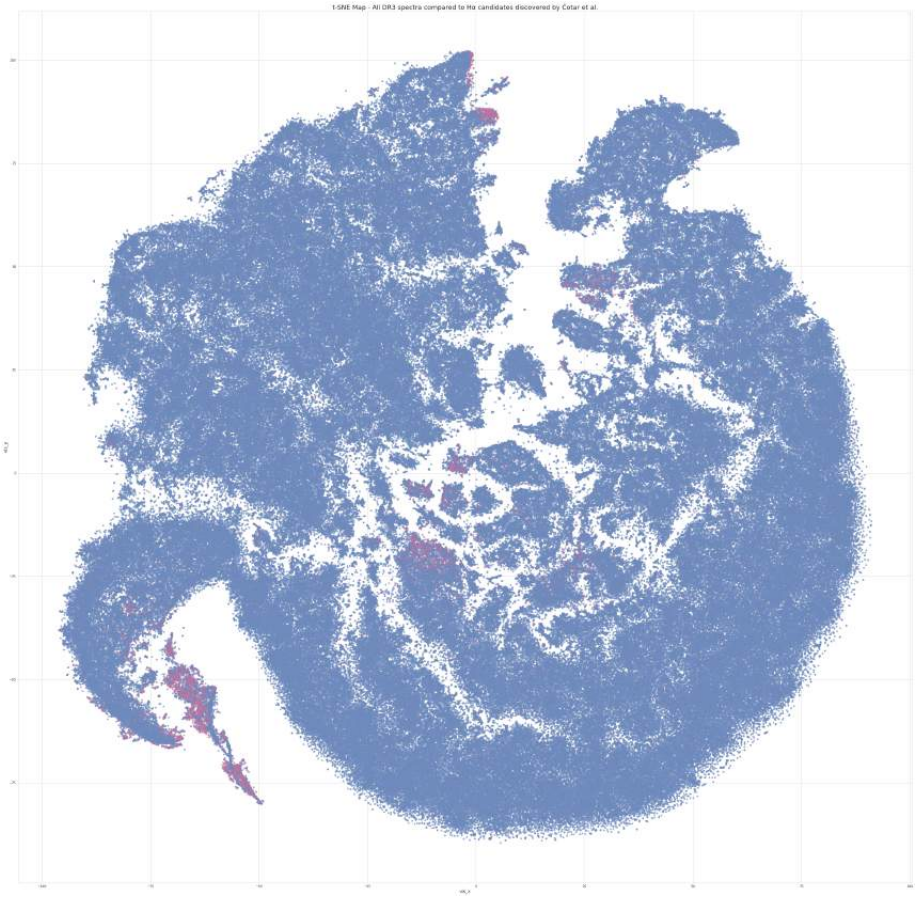

FIGURE 5.3: t-SNE map for all spectra in DR3 with Hα emission-line spectra identified by Čotar et al. tagged in pink.

5.3 Can t-SNE Be Used to Classify P Cygni and Inverse P Cygni Spectra?

Traven et al. proposed that a second application of t-SNE on Hα emission-line spectra could

cluster and classify P Cygni and inverse P Cygni emission-line spectra. This approach was tested on the Hα emission-line spectra identified by Čotar et al.. The steps are as follows:

1. Generate a t-SNE map of the Hα emission-line spectra identified by Čotar et al. using the hyper parameters suggested by Traven et al..

2. Using object IDs, map the ten clusters identified by the method proposed in Chapter 4 onto this map.

3. Validate if t-SNE is able to meaningfully separate the P Cygni and inverse P Cygni spectra identified by the method proposed in Chapter 4.

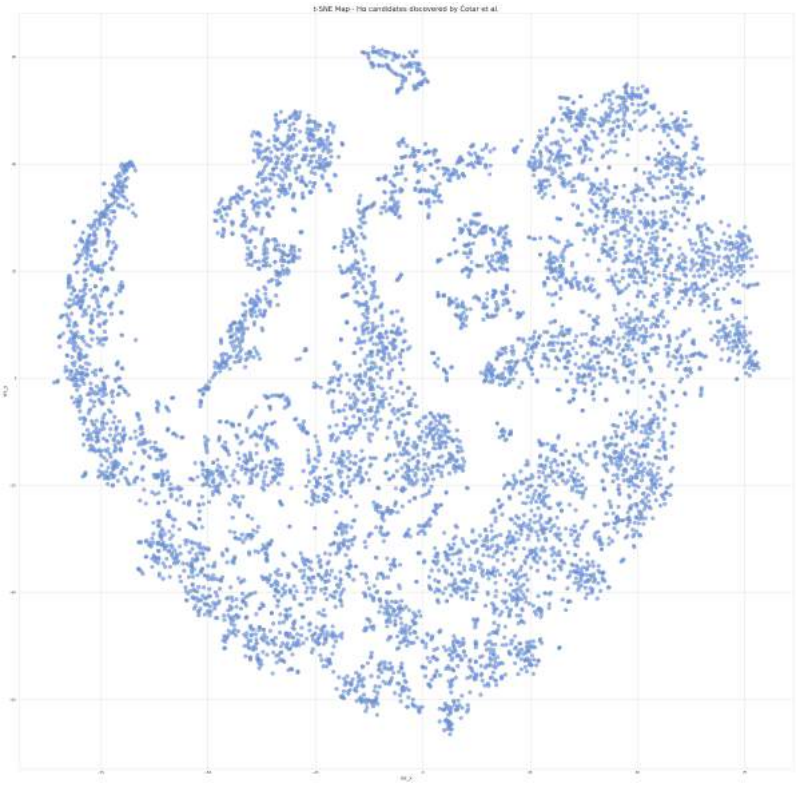

FIGURE 5.4: t-SNE map for Hα emission-line spectra identified by Čotar et al..

Note that the representation shown in Figure 5.4 is significantly simpler compared to the prior application of t-SNE to DR3. In Chapter 4, DTW and agglomerative hierarchical clustering identified ten clusters. The t-SNE map is coloured according to these clusters in Figure 5.5.

In Figure 5.5, the P Cygni emission-line spectra are indicated by the colour olive, while the inverse P Cygni spectra are represented by the colour cyan. Note that the olive points

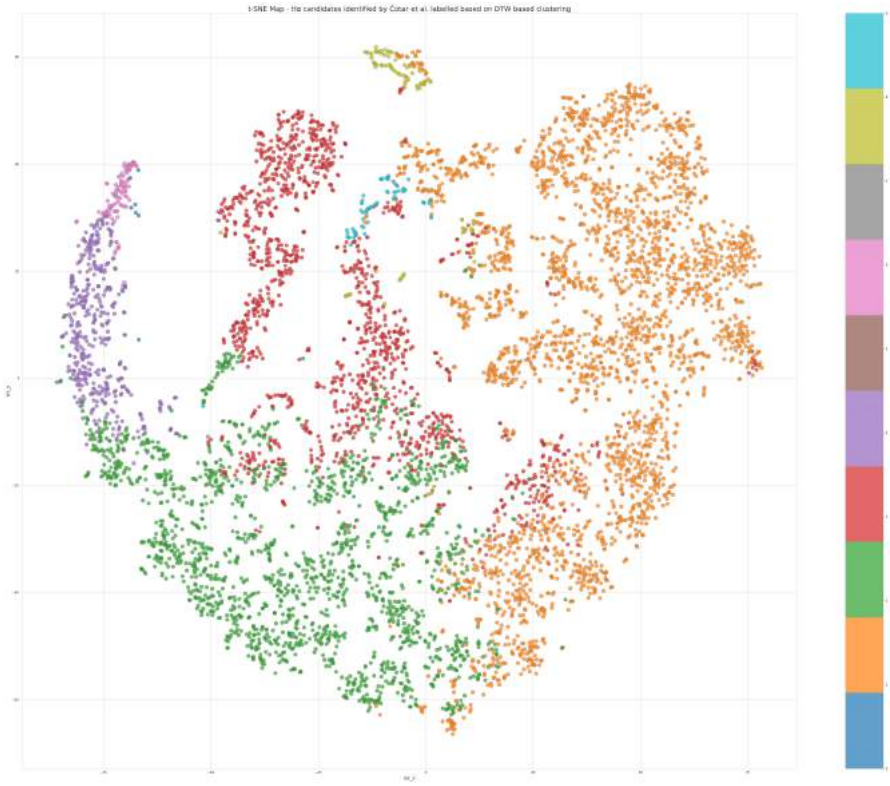

FIGURE 5.5: t-SNE map for Hα emission-line spectra identified by Čotar et al. colour-coded according to DTW-based clustering.

appear in two physically separate regions of the t-SNE map, as do the cyan points. Thus it is clear that the application of t-SNE on the Hα emission-line spectra identified by Čotar et al. does not clearly separate P Cygni and inverse P Cygni spectra into distinct regions of the t-SNE map. Hence for morphology-based classification, the DTW-based method outlined in Chapter 4 outperforms the t-SNE-based classification.

5.3.1 Examination of the P Cygni Cluster

A closer examination of the P Cygni clusters in Figure 5.6 reveals two regions on the t-SNE map that are sufficiently far apart. The P Cygni emission-line spectra in these regions were examined in further detail in order to determine which feature (or features) the t-SNE process was selecting for this distinct region separation. Splitting the regions along the y axis into two groups, $0 < y < 40$ and $40 < y < 70$, the spectra from each region were plotted on the same wavelength grid. The P Cygni cluster (Figure 5.6), split along these regions is provided in Figures 5.8 and 5.9 respectively.

The results indicate that while t-SNE may not be able to separate Hα emission-line spectra

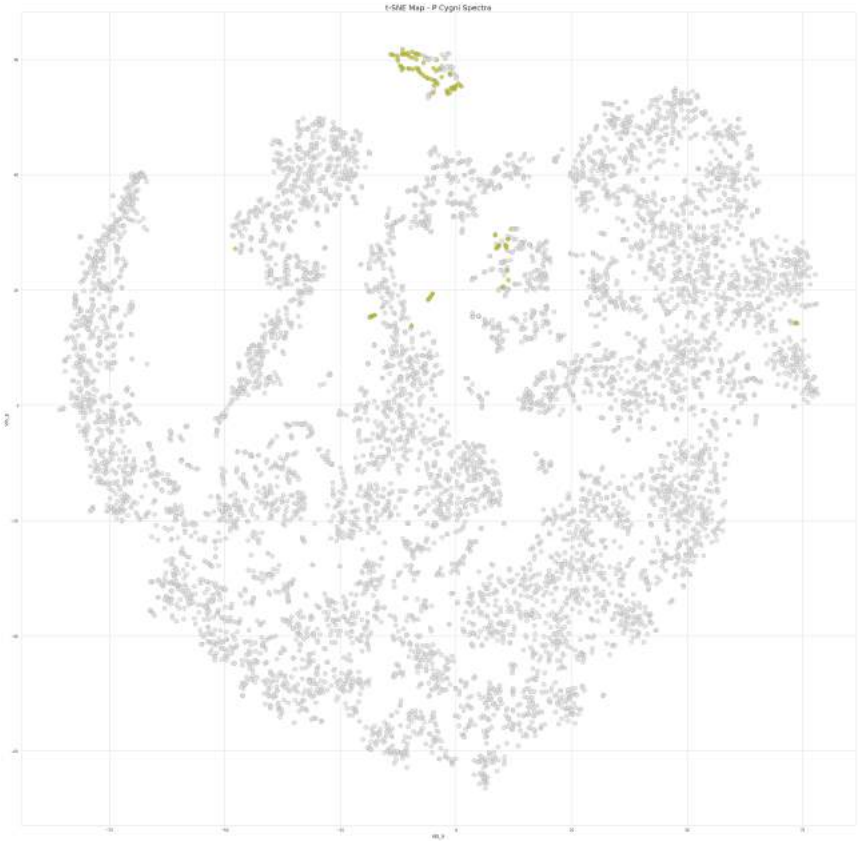

FIGURE 5.6: t-SNE map highlighting the P Cygni cluster.

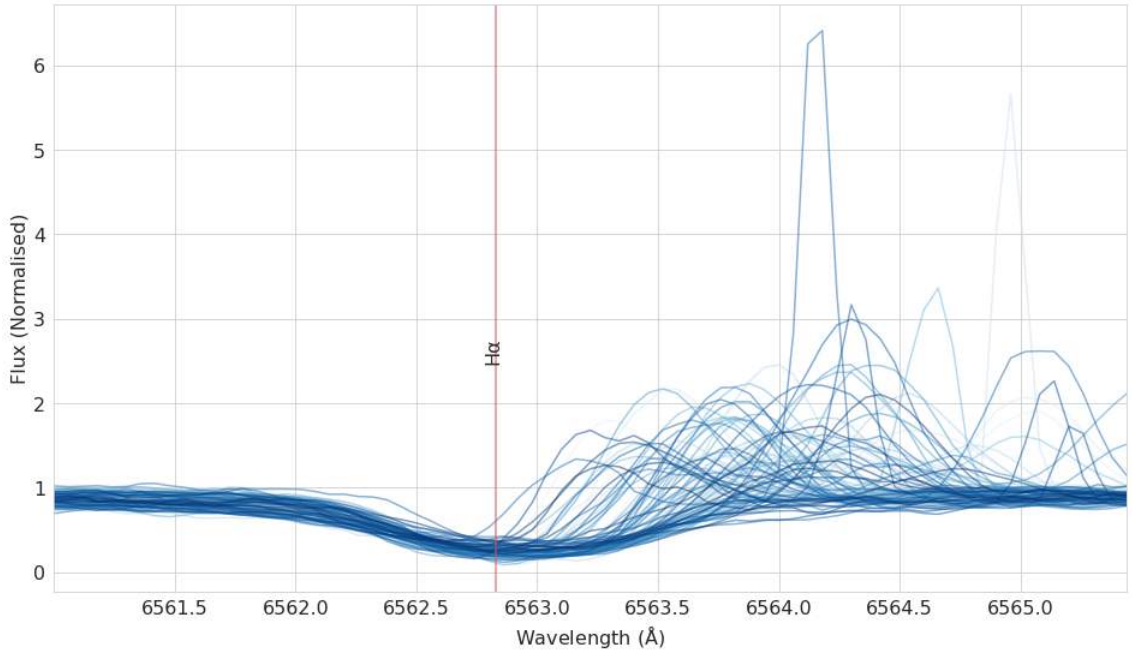

FIGURE 5.7: P Cygni spectra identified using DTW-based clustering in the Hα emission-line spectra data set released by Čotar et al.

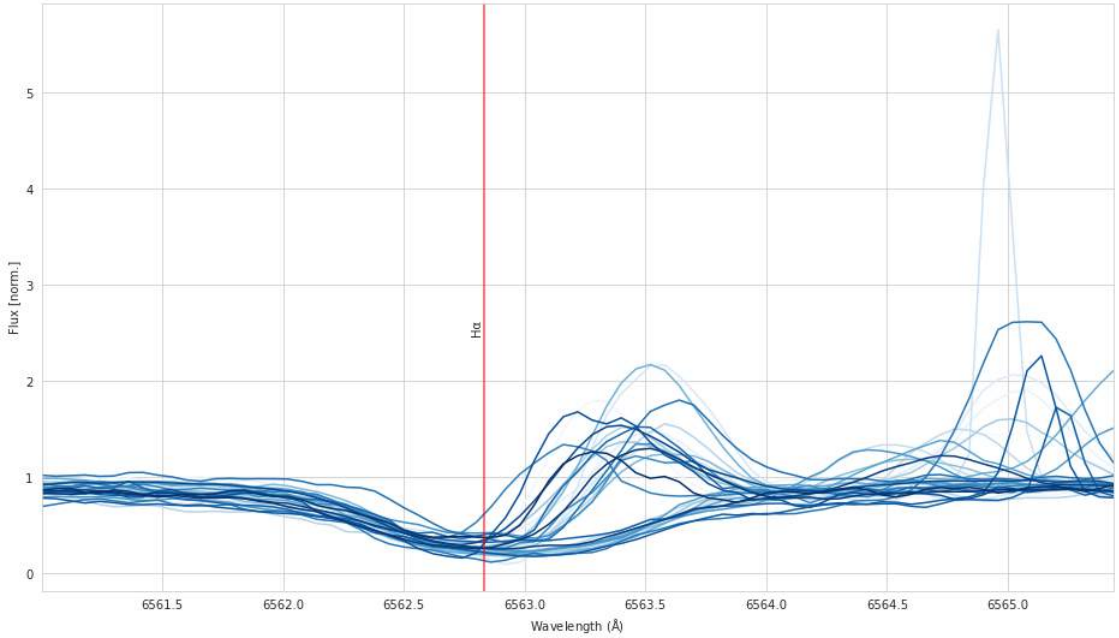

FIGURE 5.8: P Cygni spectra where t-SNE region is 0<y<40.

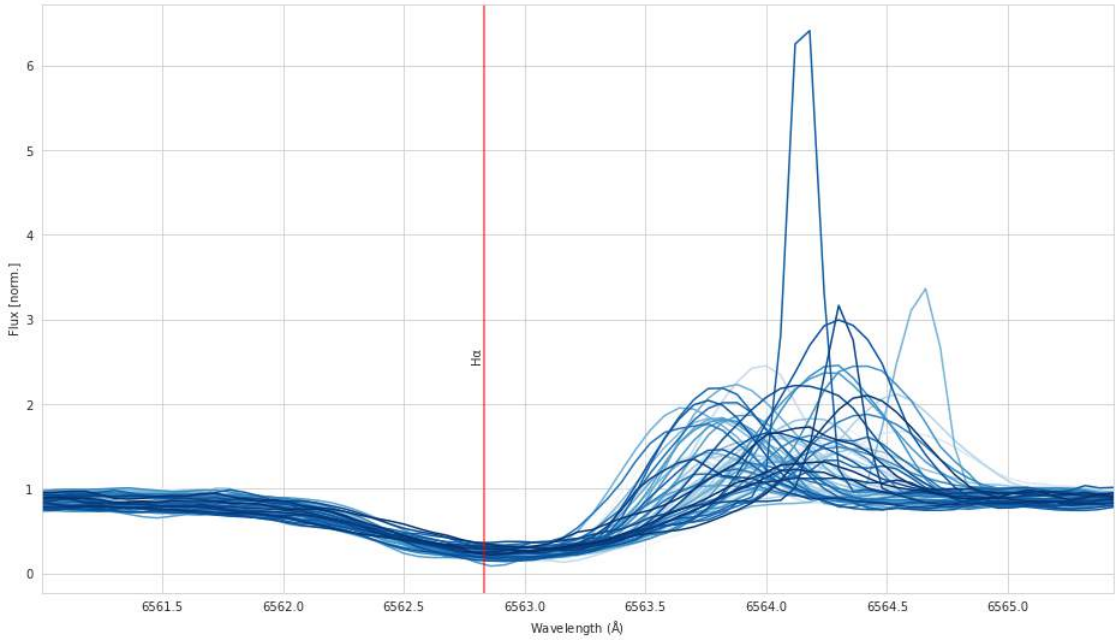

FIGURE 5.9: P Cygni spectra where t-SNE region is 40<y<70.

based on morphology as effectively as DTW based clustering, the two dimensional t-SNE space is able to discriminate based on features such as the wavelength at which the maximum flux is observed. This behaviour can be important for future feature engineering when reducing higher dimensional data to two dimensional maps such as t-SNE. One potential hypothesis is that a two dimensional representation of a high resolution and higher dimensional

spectrum introduces significant compression, such that the features that contribute to the morphology of P Cygni and inverse P Cygni spectra are no longer meaningfully represented after dimensional reduction.

5.4 Concluding remarks

t-SNE has proved extremely useful in the classification of high resolution spectra from the GALAH Survey. While some progress has been made in using this method to cluster and classify Hα emission-line spectra (Traven et al. 2017), it appears that more fine-tuned methods such as those discussed in Čotar et al. perform better on this task.

For clustering and classification of P Cygni and inverse P Cygni spectra, it appears that t-SNE may not be sufficiently sensitive to the morphology of these spectra and thus performs sub-optimally in grouping P Cygni spectra and inverse P Cygni spectra into distinct clusters on the t-SNE map. The preservation of features from high dimensions to two dimensions, such that Hα emission-line spectra can be meaningfully separated from typical spectra is not guaranteed. The pipeline for P Cygni and inverse P Cygni detection presented in Chapter 4 relies on pre-selection of Hα emission-line spectra from the GALAH survey. Thus t-SNE may not be a sufficiently robust method as a pre-processing step.

A potential hypothesis is that some higher dimensional information such as the peak wavelength (wavelength at which maximum flux occurs) has been preserved in the lower dimensional representation of P Cygni spectra and thus can perform a minor discriminatory role in the sub-classification of P Cygni spectra based on the location of the maximum flux. This may prove useful in future work, particularly in the context of feature engineering when using other machine learning methods. Thus t-SNE could be a robust method for applications and problems where maximum (or minimum) flux plays a more important role than the morphology of the spectrum. Consequently, this work will use an alternative method for this pre-processing step, and will be presented in Chapter 6.

6

Identifying and Classifying Emission-line Spectra: The End-to-End Pipeline

6.1 Drawing Conclusions from Prior Chapters

The following conclusions can be drawn from the results presented in the preceding chapters:

1. The curse of dimensionality presents a significant challenge when working with high-resolution data from million star surveys such as GALAH DR3.

2. Dimensionality reduction methods such as t-SNE may not be sufficiently robust at identifying emission-line spectra. Furthermore, the two dimensional t-SNE representation may not be sensitive to the morphological differences between emission-line spectra and non-emission line spectra.

3. Given a data set with emission-line spectra, DTW-based agglomerative hierarchical clustering can be used effectively to identify and categorise P Cygni, inverse P Cygni

and other emission-line spectra.

This chapter builds on these conclusions and presents a proof of concept pipeline for the identification and classification of emission-line spectra. DTW-based agglomerative hierarchical clustering, which was introduced in Chapter 4, forms the basis of this pipeline. Furthermore, this pipeline relies on the pre-selection of Hα emission-line spectra using an autoencoder first introduced by Čotar et al. (2021). In this chapter, the full GALAH DR3 data set with approximately 600,000 spectra will be utilised, as opposed to the much smaller Hα emission-line star data set provided by Čotar et al., which was used extensively in Chapter 4 and Chapter 5.

6.2 Identifying Hα Emission-line Spectra

Chapter 2 presented an autoencoder neural network developed by Čotar et al. that is capable of identifying Hα emission-line spectra. This method was adapted as a pre-processing step prior to the application of the DTW-based agglomerative hierarchical clustering technique demonstrated in Chapter 4. The autoencoder method is as follows:

The autoencoder learns the latent space representation of non-emission line spectra but does not learn the latent space representation of emission-line spectra accurately. This latent space represents each higher dimensional high resolution spectrum as a five dimensional vector. This process is known as encoding. Once the training phase is completed, the autoencoder is fed the total DR3 data set to generate predictions for each spectrum. This process is known as decoding.

The training data are deliberately biased towards non-emission line spectra (so-called normal or typical spectra). Thus the predicted spectra will match the data for non-emission-line spectra more accurately than for emission-line spectra. The flux difference between a predicted spectrum and its observed (original) data counterpart can be used to flag emission-line stars. This flagging can be accomplished by computing the equivalent width of this ”difference spectrum” around the Hα line.

6.2.1 The Autoencoder Architecture and Training

The autoencoder architecture presented here is similar to the work of Čotar et al. and maps high resolution spectra of dimension 4,459 to a five-dimensional latent space. The popular

deep-learning frameworks `tensorflow` (Abadi et al. 2015) and `keras` (Chollet et al. 2015) were used to develop the autoencoder. Each successive dense layer in the neural network reduces the dimensionality of the input layer successively by 75%, 50%, 25% and finally by 10%. This is visually represented in the architecture diagram in Figure 6.1.

Neural networks require non-linear activation functions in order to transform a summed, weighted input to a given node, such that the node is able to produce the so-called activation of the node (the output). This activation then serves as the input to the next node. For this activation process, non-linear functions are preferred, as they enable a neural network to learn more complex structures in comparison to neural networks that use only linear activation functions (Goodfellow et al. 2017). The ReLU function is a popular non-linear activation function that has been used extensively in training deep neural networks. ReLU is a piecewise linear function that will output the input directly if it is positive, but otherwise output zero. Most importantly, it overcomes the so-called "vanishing gradient problem". In this scenario, layers of the neural network fail to receive useful gradient information. The errors that are back-propagated through the network decrease dramatically and prevent the neural network from learning effectively. Additionally, implementing ReLU activation can reduce computational complexity (Goodfellow et al. 2017).

Further improvements have been made to this ReLU regime. Notably, with the introduction of the Parametric Rectified Linear Unit (PReLU), significant improvements have been made in the cost reduction of training when using ReLU (He et al. 2015). Given these improvements, Čotar et al. ensured that each layer is activated by the non-linear PReLU function. This work adopts the same strategy.

The training loss is minimised using the Adam optimiser, which performs gradient descent (Kingma & Ba 2014). The advantages of using Adam over stochastic gradient descent include computational efficiency and low memory use. Empirical demonstrations have indicated that Adam is particularly suited for problems with large datasets and parameters, which justifies the decisions made by Čotar et al.

Čotar et al. recommended inverting the flux values (1 - normalised flux) prior to training and prediction. Greater training stability was achieved by this inversion. Furthermore, they recommended an epoch size of 350 and a batch size of 40,000 spectra. This implies that the neural network is updated once every 40,000 spectra, while the network passes through the training dataset 350 times. For validation, 10% of the samples were selected and set

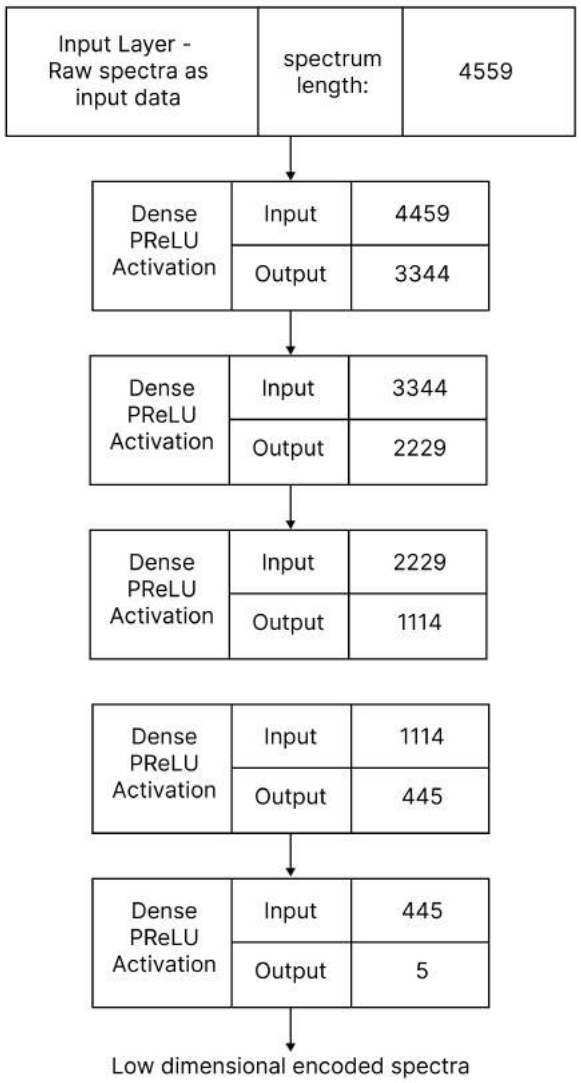

FIGURE 6.1: Visual representation of the encoder. The value in the right most column of each layer indicates the number of input and output connections to neighboring layers.

aside. This work followed the same conventions. Exploring the sensitivity of the method to these initial conditions would require an extensive investigation, one beyond the scope of this thesis.

For training, the autoencoder required a training set that is either significantly biased towards non-emission line spectra around Hα, or better yet, contains exclusively non-emission-line spectra. In order to select these spectra in DR3, the quality criteria recommended by Čotar et al. (2021), Buder et al. (2021) and Kos et al. (2017) were applied. While these criteria may not fully guarantee that the autoencoder will be trained exclusively on non-emission line spectra, prior work by Čotar et al. indicated that the criteria are sufficiently robust for the purpose of training this specific network. The Data Central SQL/ADQL catalogue query service

was used to retrieve GALAH DR3 `sobject_id` values that matched these criteria—namely:

```
SELECT sobject_id
FROM   galah_dr3.main_star
WHERE  snr_c3_iraf > 30
       AND red_flag = 0
       AND flag_sp < 16
```

Criterion	**Rationale**
SNR >30	Spectra have reduced noise contamination
`red_flag` = 0	Select spectra that have no know reduction issues
`flag_sp` <16	Do not include known emission-line spectra identified by t-SNE

TABLE 6.1: GALAH DR3 selection criteria for non-emission line spectra for training purposes.

This query returned 396,338 spectra. The red arm data was re-sampled to a common wavelength grid using the method outlined in Chapter 3. The normalised flux was then inverted and used as the training data set.

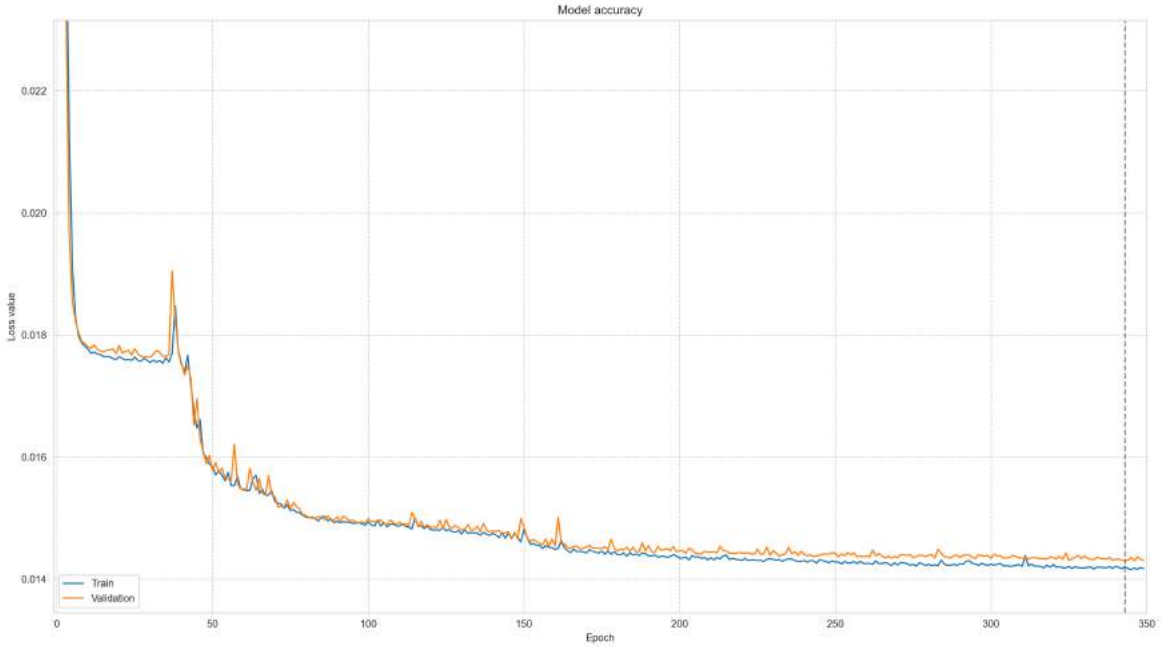

FIGURE 6.2: Prediction accuracy of the red arm training data set at different training epochs.

The autoencoder was applied to these data and the prediction error for training was computed as a sum of all absolute differences between the input and output data sets. The

results are presented in Figure 6.2, which shows the prediction error as a function of epoch. These results are quite comparable to those presented in Čotar et al..

It is expected that the autoencoder will perform poorly when attempting to reconstruct or predict emission-line spectra/fluxes. Conversely, it is expected that it will perform well on non emission-line spectra. This behaviour has significant implications for the detection of Hα emission-line spectra. This is discussed in the next section.

6.2.2 Using Difference Spectra to Identify Emission-line Spectra

Once the autoencoder was trained, data from GALAH DR3 were fed to the network. The network decoded these spectra and provided predictions for each spectrum based on the latent space representation it had learned. Since the input data during the training phase were inverted, for consistency of operation all DR3 spectra were inverted prior to being fed into the network. Predicted results from the autoencoder were also inverted.

The difference spectra between the predicted and the original DR3 spectra (observed spectra) were computed using Formula 6.1. Presented below are the inverted difference spectra for a known non emission-line spectrum (Figure 6.3) and a known emission-line spectrum (Figure 6.4) in DR3 around Hα. The emission-line spectrum, which is a P Cygni, was identified from the work presented in Chapter 4. Note that the non emission-line spectrum produces a difference spectrum with an approximately flat response, while the emission-line spectrum does not.

$$f_{difference} = f_{observed} - f_{predicted} \tag{6.1}$$

The equivalent width within the range 6561Å - 6565Å was determined using the popular Python packages `astropy` (Astropy Collaboration et al. 2013, 2018) and `specutils` (Astropy-Specutils Development Team 2019). For consistency, these widths were calculated for the inverted difference spectra. Čotar et al. used an equivalent width cut-off of 0.25 to separate emission-line spectra from DR3. The sensitivity of this cut-off parameter was tested. Selecting a lower threshold for this parameter, such as EW > 0.20, allows for a wider selection of potential Hα emission-line spectra, while a higher threshold such as EW > 0.50 may restrict the selection to only the strongest emitters. The rationale for choosing EW > 0.25 in Čotar et al. was presumably based on a trial and error approach. A similar empirical strategy to fine-tune this parameter and select a reasonable population of Hα emission-line

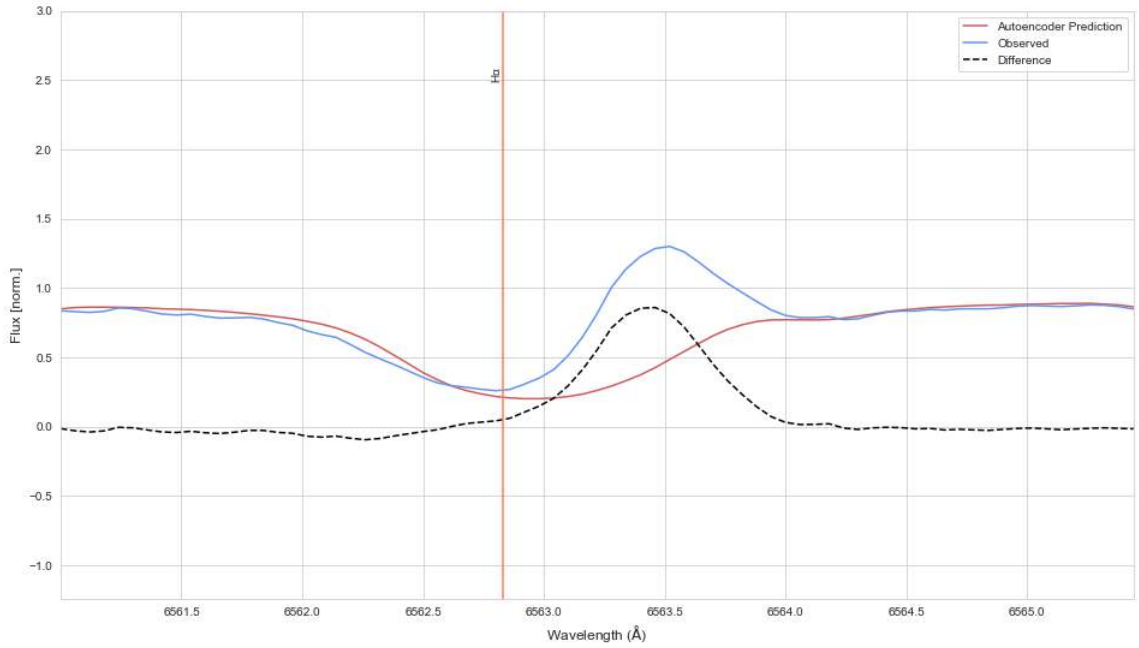

FIGURE 6.3: An emission-line spectrum (P Cygni), the autoencoder prediction and corresponding difference spectrum. Note the non-flat response of the difference spectrum. This response can be quantified by calculating its equivalent width.

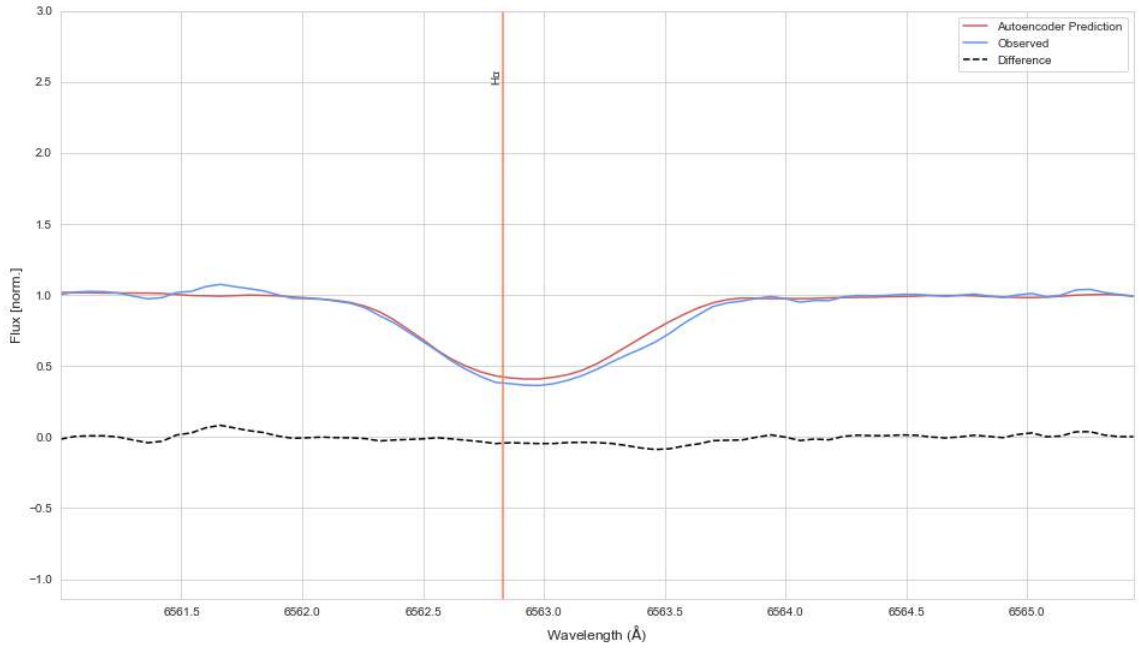

FIGURE 6.4: A non-emission-line spectrum, the autoencoder prediction and corresponding difference spectrum. Note the flat response of the difference spectrum. This response can be quantified by calculating its equivalent width.

spectra resulted in setting EW > 0.22. This value provided the best overall performance. However, this cut-off does not guarantee that the maximum number of Hα emission-line spectra was selected.

Given that Čotar et al. used a different population for their work, the Hα emission-line spectra discovered cannot be directly compared to those found in GALAH DR3 using the method above. It can be noted that due to the inclusion of data from other surveys, only 4,556 out of 10,364 objects identified by Čotar et al. were present in the GALAH DR3 sample, which comprised 396,338 spectra. These 4,556 were also captured and recovered within the Hα emission-line spectra discovered using the autoencoder and setting EW > 0.22 (Figure 6.5).

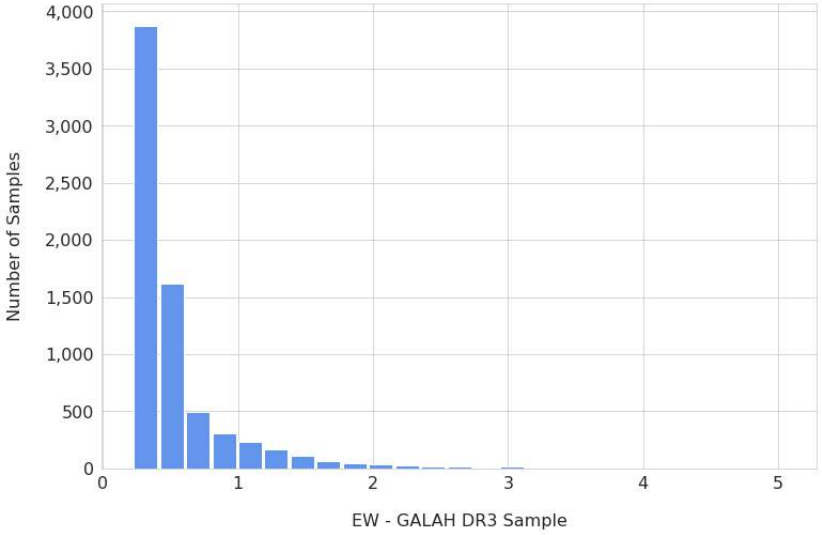

FIGURE 6.5: The equivalent width (EW) distribution of the inverted difference spectra of the emission-line spectra identified in GALAH DR3. Here EW > 0.22.

6.3 Applying Dynamic Time Warping and Agglomerative Hierarchical Clustering

Of the 396,338 spectra that were fed to the decoder, 7,067 were identified as Hα emission-line spectra. Once the Hα emission-line spectra were selected, the data were passed to the next step of the analytics pipeline. The DTW distance matrix was computed for each spectrum, ensuring that only masked data were presented to the `FastDTW` algorithm (the masked region being 6561Å - 6565Å; Traven et al. 2017). As mentioned in Chapter 4, this approach

significantly reduces run-time and computational complexity. The results are presented in Figures 6.6 and 6.7.

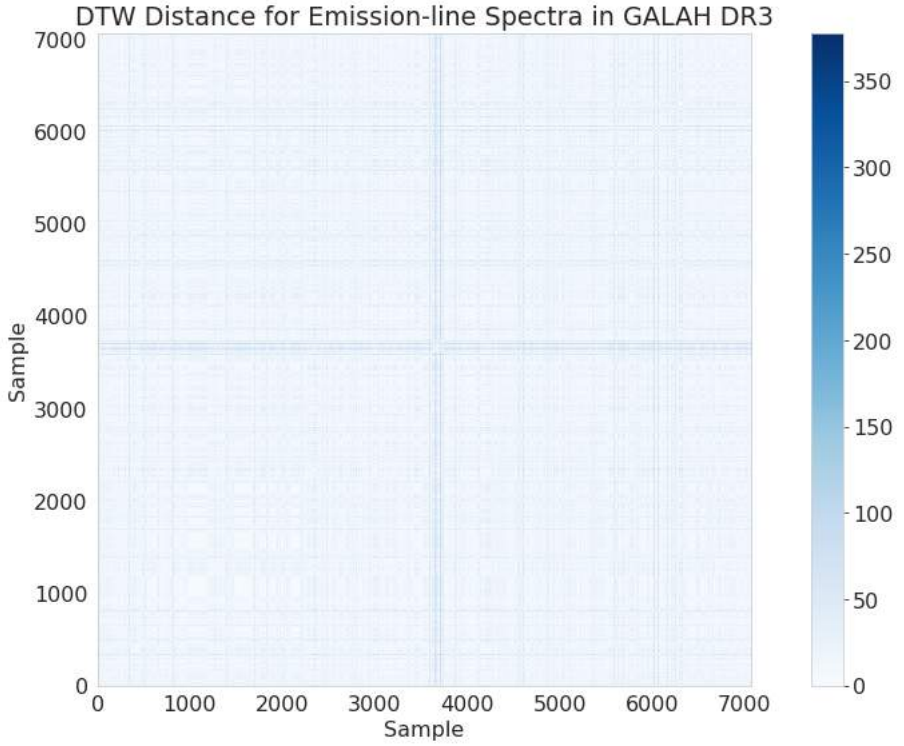

FIGURE 6.6: Pairwise DTW distances for emission line spectra identified in DR3. Darker colours indicate samples that are dissimilar.

Once the distance calculation was completed, agglomerative hierarchical clustering was performed ensuring complete linkage. At this stage, this work deviated from the number of clusters used in Chapter 4, primarily because many of the Hα emission-line spectra identified in the autoencoder stage of the pipeline exhibited noticeably more complex emission-line morphologies than those identified by Čotar et al..

This richer variety of morphologies can be attributed to the fact that the DR3 population is fundamentally different to the survey data used by Čotar et al.. This can be demonstrated by observing the number of P Cygni spectra that separate into a cluster as the number of clusters is increased beyond the value (10) used in Chapter 4. With Hα emission-line stars identified in the DR3 dataset, the maximum number of P Cygni and inverse P Cygni were separated when the number of clusters was set to 45. This is significantly higher than what was observed in the results presented in Chapter 4. The tree generated by agglomerative hierarchical clustering is thus cut at a depth of 45 (as opposed to 10) for this specific sample for a maximal separation of P Cygni and inverse P Cygni spectra.

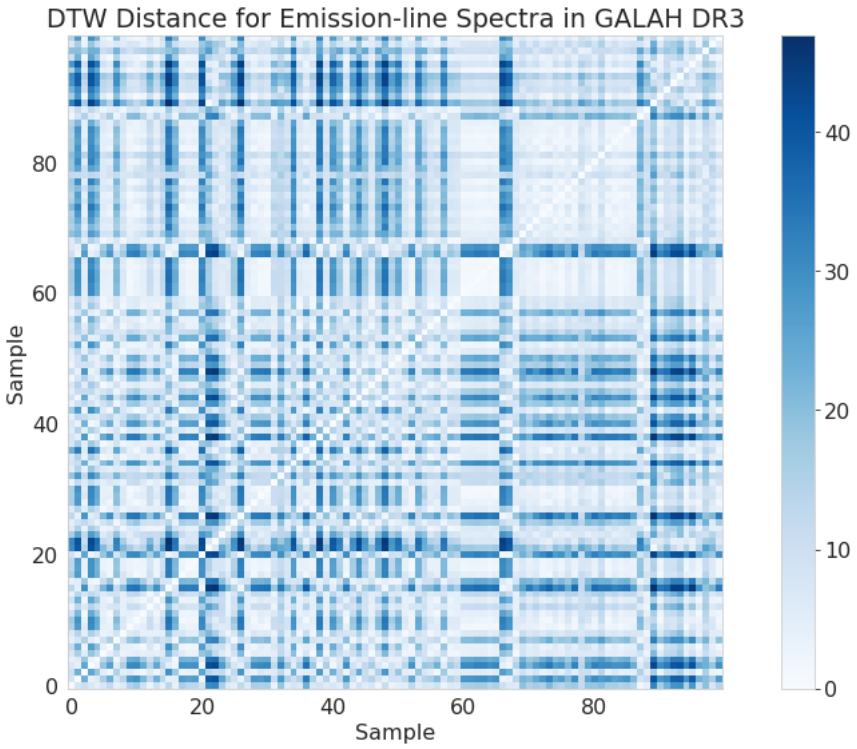

FIGURE 6.7: Pairwise DTW distances for emission line spectra identified in DR3 (zoomed). Darker colours indicate samples that are dissimilar.

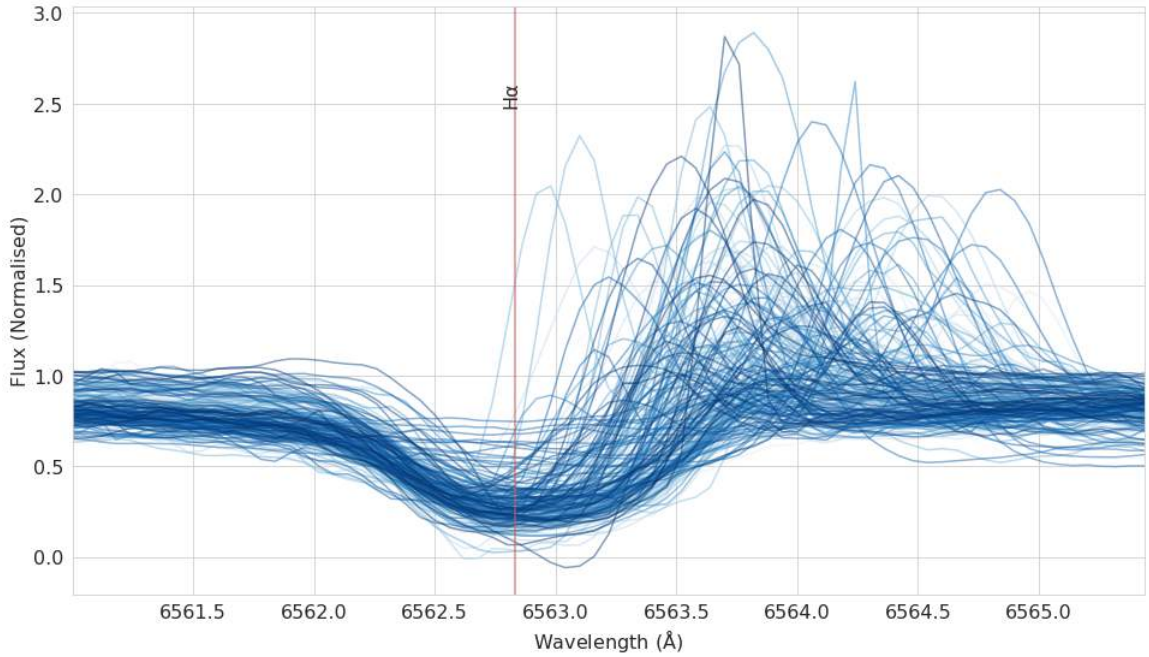

FIGURE 6.8: Ensemble plot of 243 P Cygni spectra identified in DR3 using DTW.

The significant advantage of utilizing a higher number of clusters than 10 is that if, as in the case of DR3, more peculiar emission-line spectra are present, they will be separated from

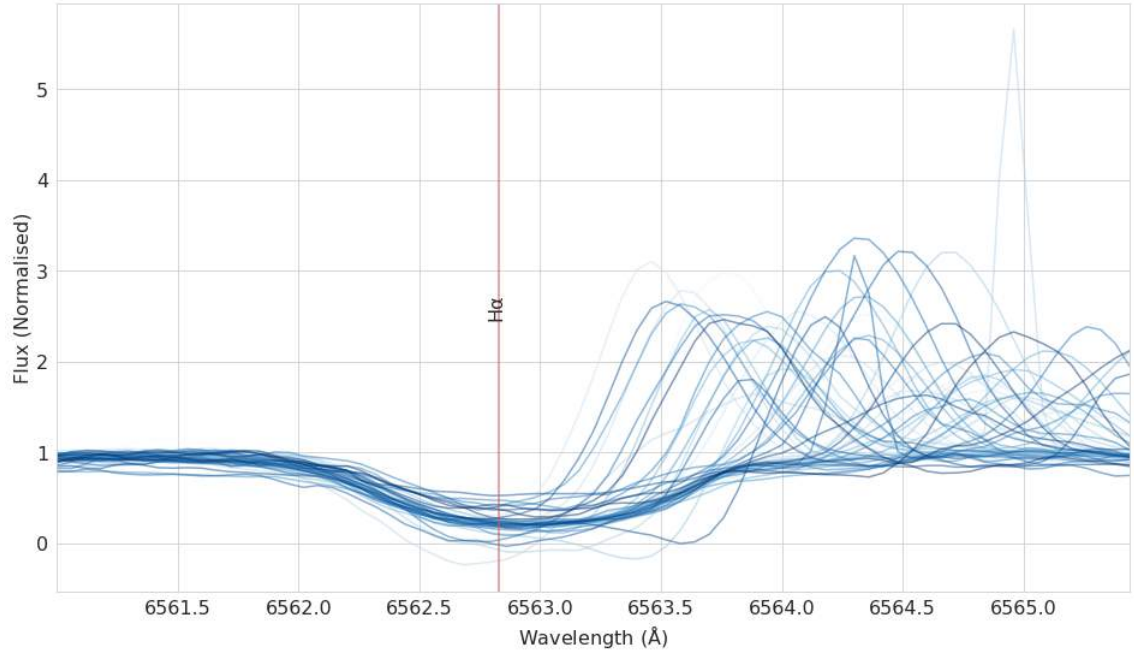

FIGURE 6.9: Ensemble plot of 53 additional P Cygni spectra identified in DR3 using DTW. These were not included in the main P Cygni cluster but appeared in a separate group, likely due to a less prominent absorption feature blueward of Hα.

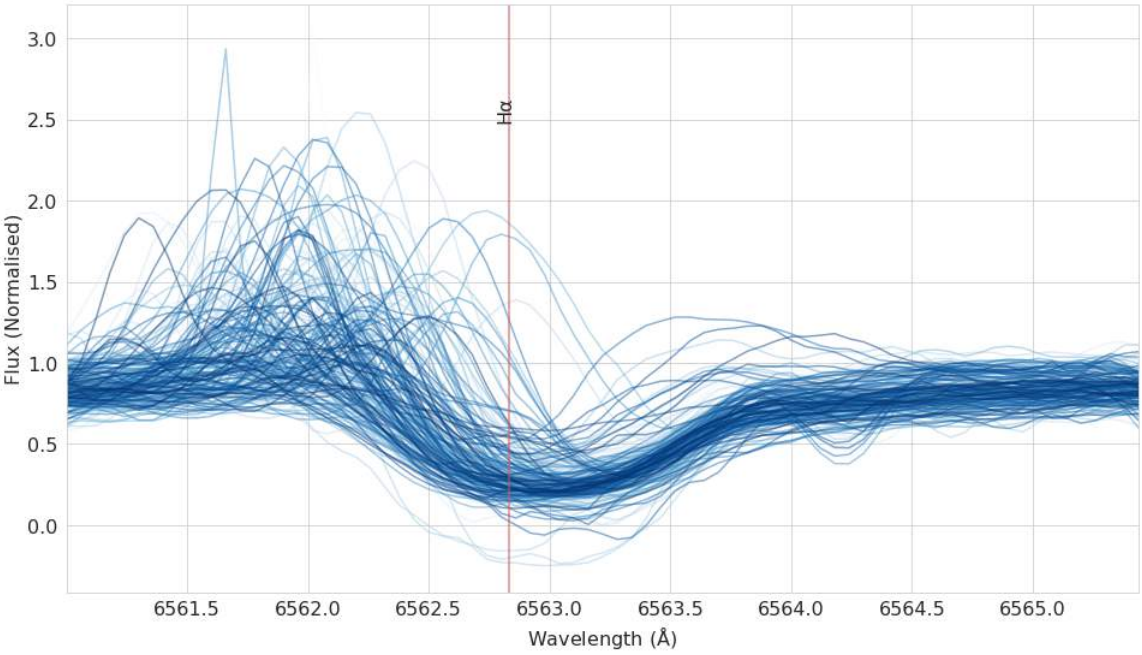

FIGURE 6.10: Ensemble plot of 219 inverse P Cygni spectra identified in DR3 using DTW.

the total data set more effectively, i.e., there is no scheme by which these extremely peculiar yet rare spectra can hide in a larger set of data points. These peculiar spectra would form their own minority clusters within the overall hierarchy of the tree generated by the clustering

procedure. As far as the author is aware, there exists no precedent in literature for selecting 45 distinct classes for emission-line spectra. However, given that the process outlined here is entirely data driven, it is justifiable since this value creates maximal separation of P Cygni and inverse P Cygni spectra, as well as the separation of peculiar sub-species in DR3 which have not been classified previously. Additional examples of some of these classes are provided in Figures 6.11, 6.12 and 6.13. In Chapter 4, the silhouette score was introduced as a means to discriminate between potential cluster sizes. However, this work did not find a clear relationship between the silhouette score and the number of P Cygni and inverse P Cygni spectra that were separated from the broader group of emission-line spectra. Further work is required to determine the exact relationship (if any) between the number of clusters, the silhouette score and the number of P Cygni and inverse P Cyni spectra. This work is currently beyond the scope of this thesis. Further inspection of key classes were carried out using an interactive plotting tool created by using the `plotly` package.

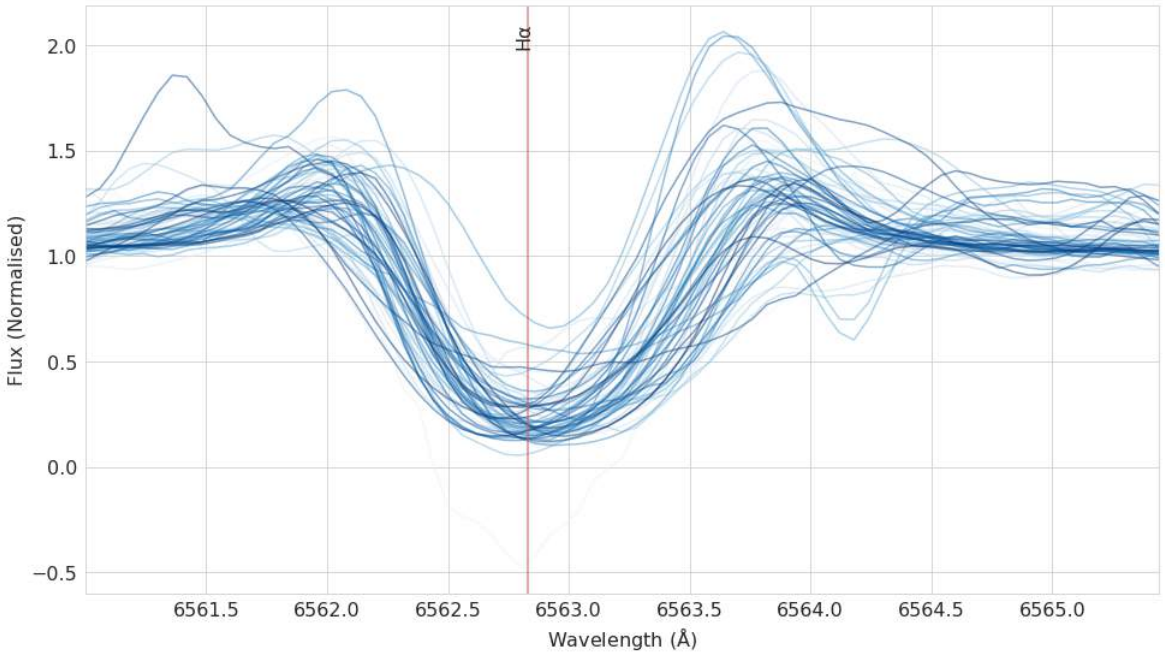

FIGURE 6.11: Ensemble plot of 67 double peaked emission-line spectra identified in DR3 using DTW.

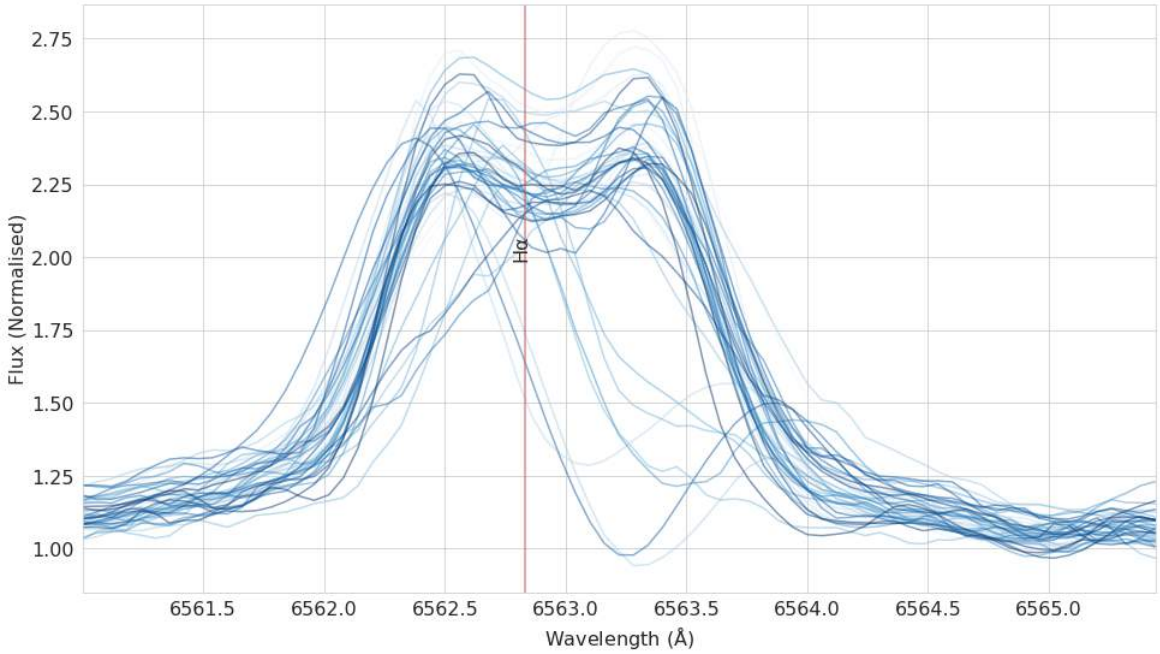

FIGURE 6.12: Ensemble plot of 46 self-absorption type spectra identified in DR3 using DTW.

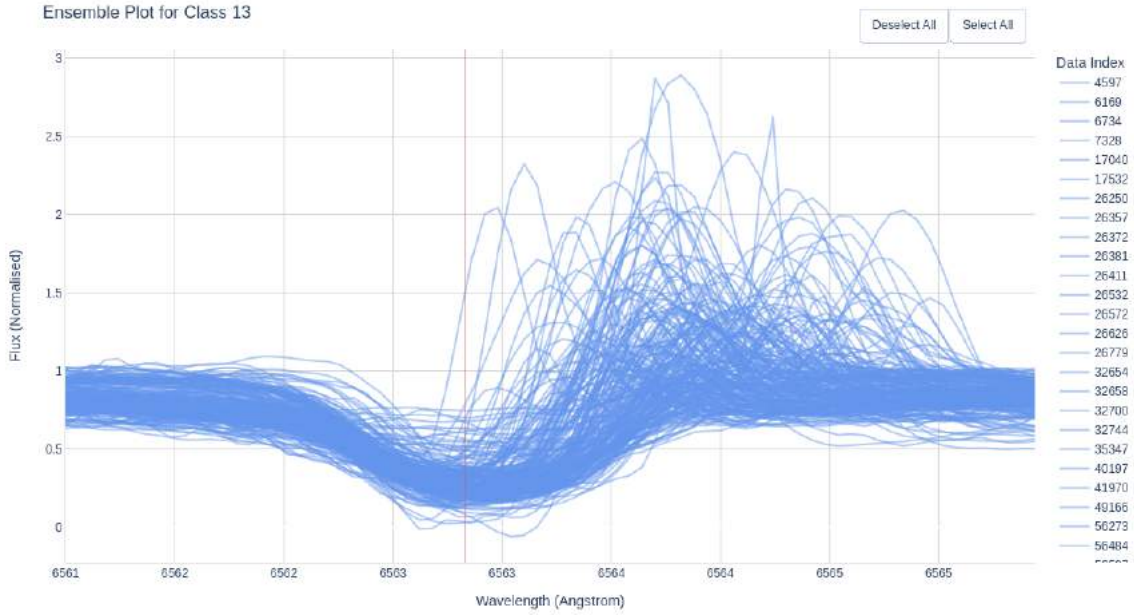

FIGURE 6.13: Inspecting a class of spectra identified by DTW using a `plotly` plot which changes interactively based on the object selected.

6.4 Concluding Remarks

Identifying atypical spectra such as emission-line spectra from a data set that contains a majority of typical spectra is a challenging task for machine learning methods. After analysing the requirement and end goals, a unique and powerful combination of methods such as dimensionality reduction, anomaly detection using autoencoders and dynamic time warping was carefully selected to identify and cluster emission-line spectra in the GALAH DR3 dataset.

Apart from selecting the number of clusters, the pipeline presented in this work requires significantly less human intervention than other methods in literature. Crucially, it does not require human intervention to manually identify emission-line spectra from a large data set such as GALAH DR3.

Using these methods, this work was able to identify 296 P Cygni spectra (Figure 6.8 and 6.9) and 219 inverse P Cygni spectra (Figure 6.10), as well as numerous other classes of stars with Hα emissions from a subset of 7,067 emission-line spectra. This leads to the conclusion that, in instances where atypical spectra are hidden in a data set which comprises a majority of typical spectra, this novel approach is able to outperform the machine learning methods presented to date in the literature. Further work is required to characterise the over 500 spectra identified as P Cygni and inverse P Cygni and will form the future direction of this work

7

Conclusions and Future Work

7.1 Conclusions

Given the deluge of spectroscopic data from modern surveys, manual methods, rule-based methods and certain machine-learning methods may not be suitable to address the twin problems of identification and classification of non-typical spectra such as Hα emission-line stars. The preceding chapters demonstrated that these challenges can be overcome using a combination of dimensionality reduction, anomaly detection and spectral morphology-based clustering that utilises dynamic time warping and agglomerative hierarchical clustering.

The methods presented in this work provide significant advantages over other recent methods that rely on manual human intervention to identify and classify P Cygni, inverse P Cygni and other species of emission-line spectra such as Zhang et al. (2021) and Zhao et al. (2012). In addition to requiring minimal human intervention when setting the machine learning parameters, a key advantage is that these methods do not require a labelled training data set, further reducing the need for human intervention. This work demonstrated that if the

code, data structures, algorithms and strategy are picked carefully, emission-line spectra can be identified and classified with reasonable computational efficiency with complexity $O(N)$, without requiring a human being to review manually thousands, hundreds of thousands or even potentially millions of spectra.

The autoencoder used in this analysis functions as an anomaly detector, capable of flagging spectra with emission-line features. This is a mature and well-established methodology in domains such as credit card fraud detection, and the approach can be extended to other areas of astronomy that require the detection of atypical signals from a large volume of more typical signals. Similarly, dynamic time warping is a well-established signal processing and machine-learning method that is sensitive to shapes and morphologies of signals over other features, and consequently can be applied to other fields in astronomy where the morphological features of a signal play a dominant role over other considerations. To this author's knowledge, casting spectra as time series and using DTW has not been used to identify and classify astronomical spectra in the literature, and consequently is a novel approach.

With agglomerative hierarchical clustering, this work demonstrated that the number of clusters and consequently the depth of the hierarchical tree is sensitive to the population mix over which the algorithm is run. In the case of GALAH DR3, the maximum numbers of P Cygni and inverse P Cygni spectra were separated out when the number of clusters was set to 45. While the literature that informed this work suggested that the number of clusters must be greater than 6 and under 10, determining this value for the emission-line population present in DR3 required an iterative approach of setting the number of clusters and tracking the number of spectra separated from the primary population as P Cygni and inverse P Cygni. While this is not particularly time consuming, it does require human intervention to define and specify a parameter to extract meaningful and useful clusters from the broader emission-line population.

The resulting larger number of clusters was useful for the classification of more atypical and exotic spectra. This "over-classification" could be extremely beneficial for researchers who study still more exotic emission-line phenomena, as well as being a useful mechanism to foster further analysis on whether some of the minority classes can indeed be combined with the majority classes that appear at higher levels of the hierarchical tree. Such decisions can be left to the researcher on the basis of what specific type of object is being studied. For example, it is clear that the clusters presented below can be combined meaningfully into

a single P Cygni cluster, thus netting a total of 296 P Cygni spectra from the DR3 sample (Figure **??** and **??**). Over-classification, prior to merging, is thus a beneficial outcome of this approach.

In identifying 7,067 emission line spectra, including over 200 P Cygni and 200 inverse P Cygni spectra, this work provides a more accurate estimation of the population of emission line stars in the DR3 survey. These spectra can now be separated from the typical spectra and processed by more finely tuned reduction pipelines that can re-evaluate and more accurately estimate their stellar parameters. This improves the overall effectiveness of the survey and helps further understanding of emission-line spectra in general.

The methodology discussed in Chapter 6 is sensitive to a number of factors, including, the type and population of stars captured by a survey. For example, if a survey is biased towards young stars or, as in the case of the Gaia ESO survey, open clusters, it is reasonable to expect a high proportion of emission-line stars in the raw data. In this scenario the researcher may have to adjust or remove the anomaly detection-based autoencoder from the pipeline suggested in Chapter 6. Autoencoders can excel at detecting anomalies in more typical-looking data, but if the data contains a majority of atypical spectra, the performance of the autoencoder must be evaluated more thoroughly. This work hypothesises that in such scenarios the performance may degrade and hence it may be beneficial to bypass or remove this step and proceed directly to computing DTW distances, followed by agglomerative hierarchical clustering.

The equivalent width cut-off that selects Hα emission-line spectra from the broader DR3 data set is another significant parameter that impacts the final population of classified emission-line stars. Čotar et al. placed this value at 0.25, but the degree to which this parameter should be tailored to the underlying sample population remains unclear; unlike the number of clusters, there exists no suitable guiding values for this parameter in the literature apart from the figure presented in Čotar et al.. This parameter may be linked to the decoding accuracy of the autoencoder, but exploring the details of this relationship would be beyond the scope of this work.

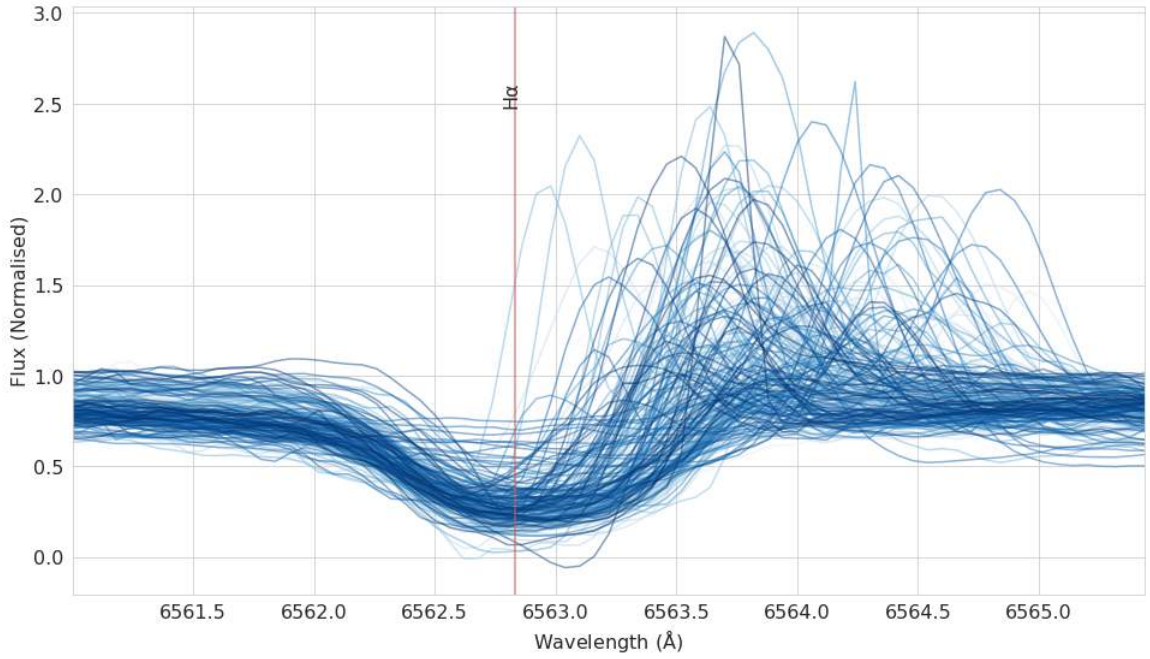

FIGURE 7.1: Ensemble plot of 243 P Cygni spectra identified in GALAH DR3 using DTW.

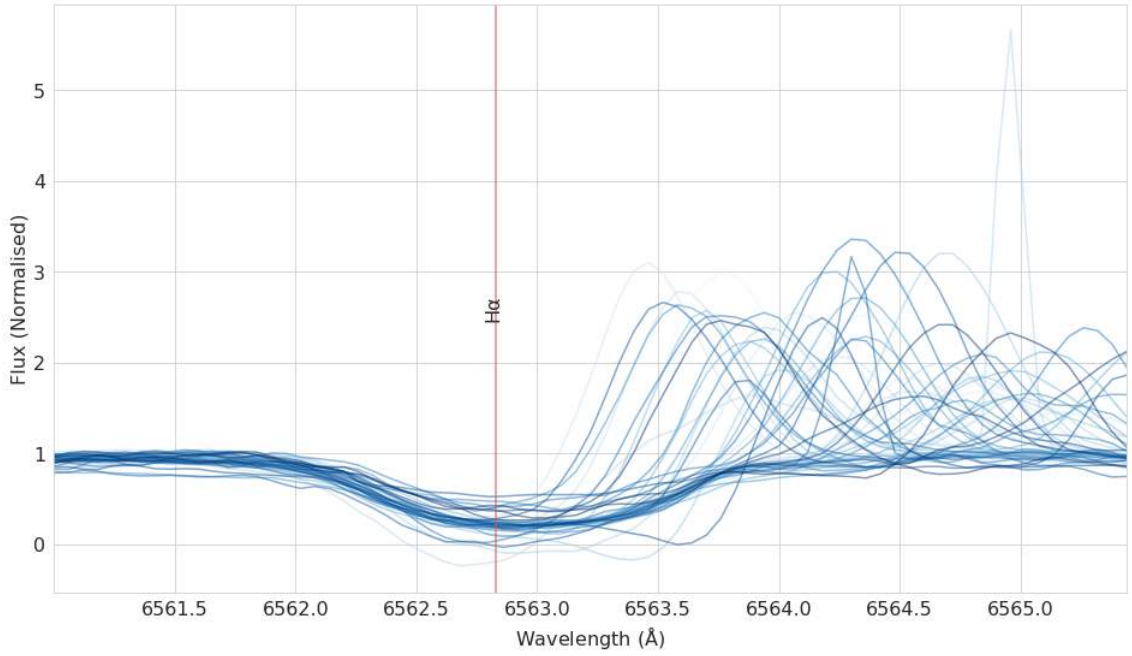

FIGURE 7.2: Ensemble plot of 53 P Cygni spectra identified in GALAH DR3 using DTW. These were not included in the main P Cygni cluster but rather appeared in a separate group, likely due to a less prominent absorption feature to the blueward of Hα. Hence, these can be combined with the majority P Cygni cluster to give 296 P Cygni spectra in total.

7.2 Future Work and Potential Applications of the Methods Discussed in This Work

This section presents several future directions and starting points that go beyond this work. Based on the results in Chapter 6, this work can be extended to identify and classify emission-line spectra in future releases of GALAH as well as surveys such as Gaia. This work also prompts further exploration of dimensionality reduction methods that are suited for the exploration of spectra from large-scale surveys.

7.2.1 Emission-line Spectra in GALAH DR4

The GALAH survey has yielded significantly more data since DR3, with over 800,000 unique targets observed to date. While this has not yet been released to the public, they are available within the GALAH collaboration, and hence can be run through the pipeline described in this work. Once the emission-line spectra in this larger sample have been identified and classified, a catalogue of emission-line stars, and in particular P Cygni and inverse P Cygni spectra can be produced, and ideally published, including cross-matches with Gaia and literature (e.g., via SIMBAD). Furthermore, a detailed analysis of this larger sample would provide an opportunity to—as mentioned in the previous section—modelling the sensitivity of the equivalent width cut-off and explore the impact of the number of clusters on classification accuracy.

7.2.2 Characterising Emission-line Spectra

Given the large number of emission-line classes, and the 296 P Cygni and 219 inverse P Cygni spectra that were identified, it would be extremely useful to characterise each group based on the type of star producing those spectra. Crucially, the stellar parameters of these members will also warrant reevaluation, as the spectra deviate significantly from more typical spectra, and so the stellar parameters for these objects determined by standard pipelines are likely to have higher uncertainties compared to those of more typical spectra. Parameters such as effective temperature and stellar masses can be reevaluated for these emission-line spectra outside the primary pipelines. For the various categories of emission-line stars, it would be useful to measure the properties of their emission spectra, for example the inflow and outflow

wind velocities of P Cygni and inverse P Cygni spectra. Expanding the binary mask that was applied to isolate the region near the Hα region could also potentially reveal objects that have high wind velocities. The methods presented in this work can be suitably adapted to study these high wind velocity objects.

7.2.3 Extension to Other Domains in Astronomy

Dynamic time (wavelength) warping and agglomerative hierarchical clustering are sufficiently generalised methods that can be adapted to a range of problems and subdomains: for example, they can be used to cluster light curves, gravitational wave signals and radio spectra. These domains face similar challenges in identifying classes of atypical signals from data sets containing more typical data or signals. Combined with an anomaly detection mechanism such as an autoencoder, these methods can be used to detect highly specific morphologies and signals hidden in large data sets containing predominantly typical (normal or non-anomalous) data.

7.2.4 Exploring Dimensionality Reduction for High Resolution Spectra

Dimensionality reduction remains an extremely useful tool when analysing high-dimensional data and feature spaces, thus serving as a potential solution to the "curse of dimensionality". However, it remains unclear which suitable lower-dimensional representation of high-resolution spectra is more useful in a particular context or set of problems. For example, in the context of clustering morphologically similar spectra, it can be argued that the 5-dimensional representation of the spectra in a latent space formed by the autoencoder was more discriminating than the 2-dimensional space formed using t-SNE.

This raises the question as to whether there are general principles or rules that can dictate these decisions for future work—and if applications where clustering is based on another parameter (e.g. effective temperature or metallicity), which lower dimensional representation will yield the best performance. As the volume and complexity of data from large scale surveys continues to grow, determining the most suitable lower-dimensional representation or projection of spectra for a given problem can be extremely useful in reducing the complexity of the data analytics process, while also reducing the computing overheads. These and related questions remain open for further investigation.

7.2.5 Building Training Data Sets for Supervised Learning

One of the key hurdles a researcher may face when attempting to identify and classify patterns in data using automated methods and machine learning is the lack of training data. Supervised machine learning methods cannot be used in such circumstances. With this work, hundreds of P Cygni and inverse P Cygni spectra have been identified, and other unusual species of emission-line spectra have also been detected. These can potentially be used to generate a training set for supervised learning algorithms. Care should be taken however when developing these training data sets. It is important that the training data set accurately reflects the statistics of the population it is attempting to generalise and represent. Thus approaches like stratified sampling must be incorporated into these efforts for the best overall results.

7.2.6 Identifying Redshifted Quasars and Galaxies in Gaia DR3 Spectroscopic Data

The Gaia DR3 data set which was released recently (Gaia-Collaboration et al. 2021) contains hundreds of millions of low resolution BP/RP spectra, which present both a challenge and a remarkable opportunity for innovative data analysis methods. One application of the methodology developed here would be the identification and classification of quasar and galaxy spectra. For quasars and galaxies across a range of redshifts (and hence wavelengths), it is expected that characteristic emission and absorption-line morphologies could be located almost anywhere in the optical spectrum. Hence a DTW-based method could potentially be used for the automated detection and classification of redshifted objects by identifying their absorption or emission features.

Modern observational astronomy is in a golden age of big data, in which major surveys and projects—yielding information on anywhere from hundreds of thousands to hundreds of millions of objects—are playing a key, if not dominant role. This brave new world calls for new techniques to exploit the wealth of data being produced both efficiently and with enough sophistication to not require intensive human interaction. The methodology developed and tested in this work has the potential to become such a technique.

A

Appendix

A.1 Appendix - Chapter 5

Chapter 5 evaluated the efficacy of t-SNE at clustering emission-line spectra. The known emission-line spectra from Čotar et al. were projected onto the same t-SNE space as the set of all GALAH DR3 spectra, in order to determine the clustering efficacy of t-SNE. The results are presented below in Figures A.1 and A.2.

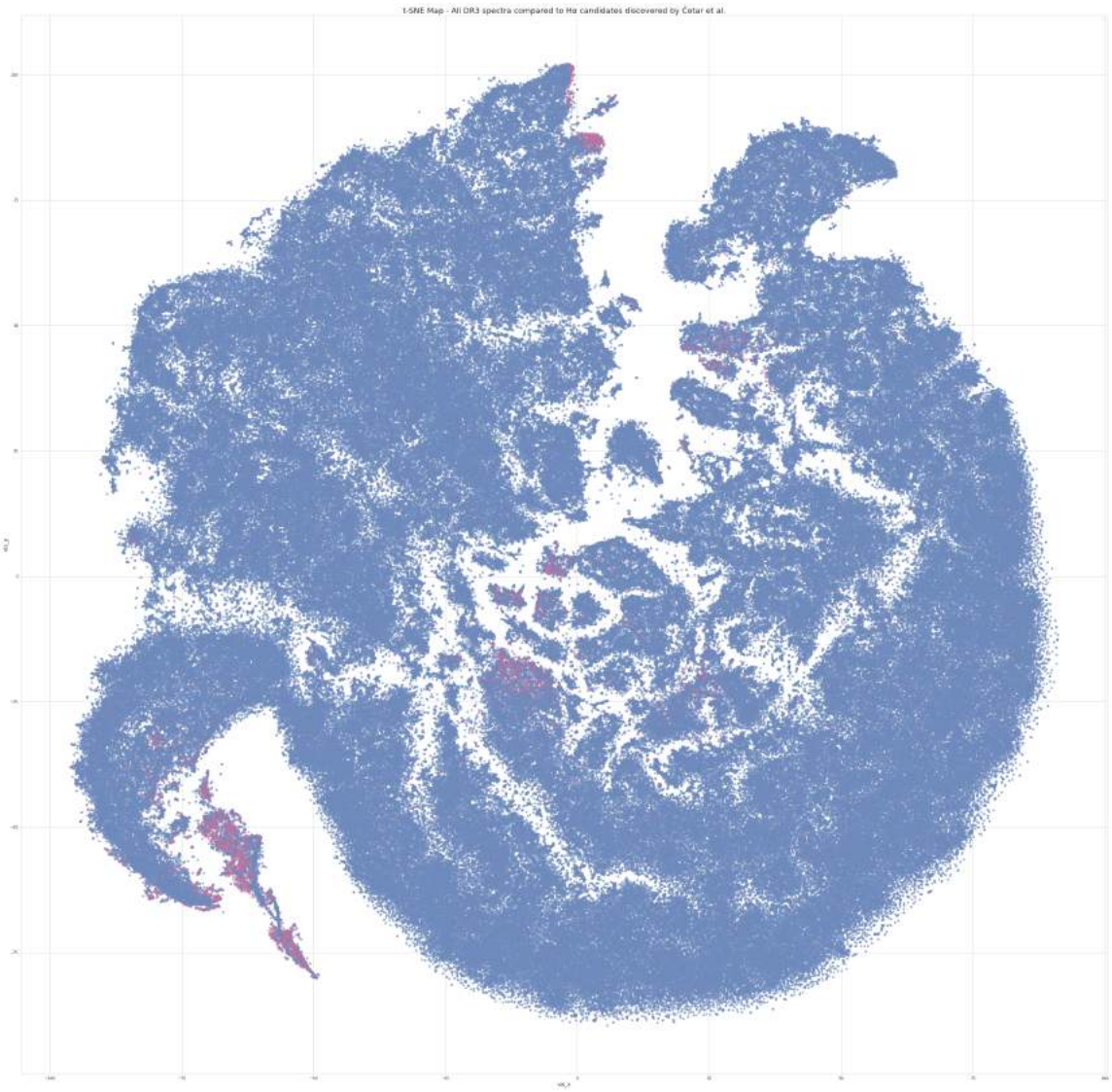

FIGURE A.1: t-SNE map for all spectra in GALAH DR3, with the Hα emission-line spectra identified by Čotar et al. (2021) indicated in pink.

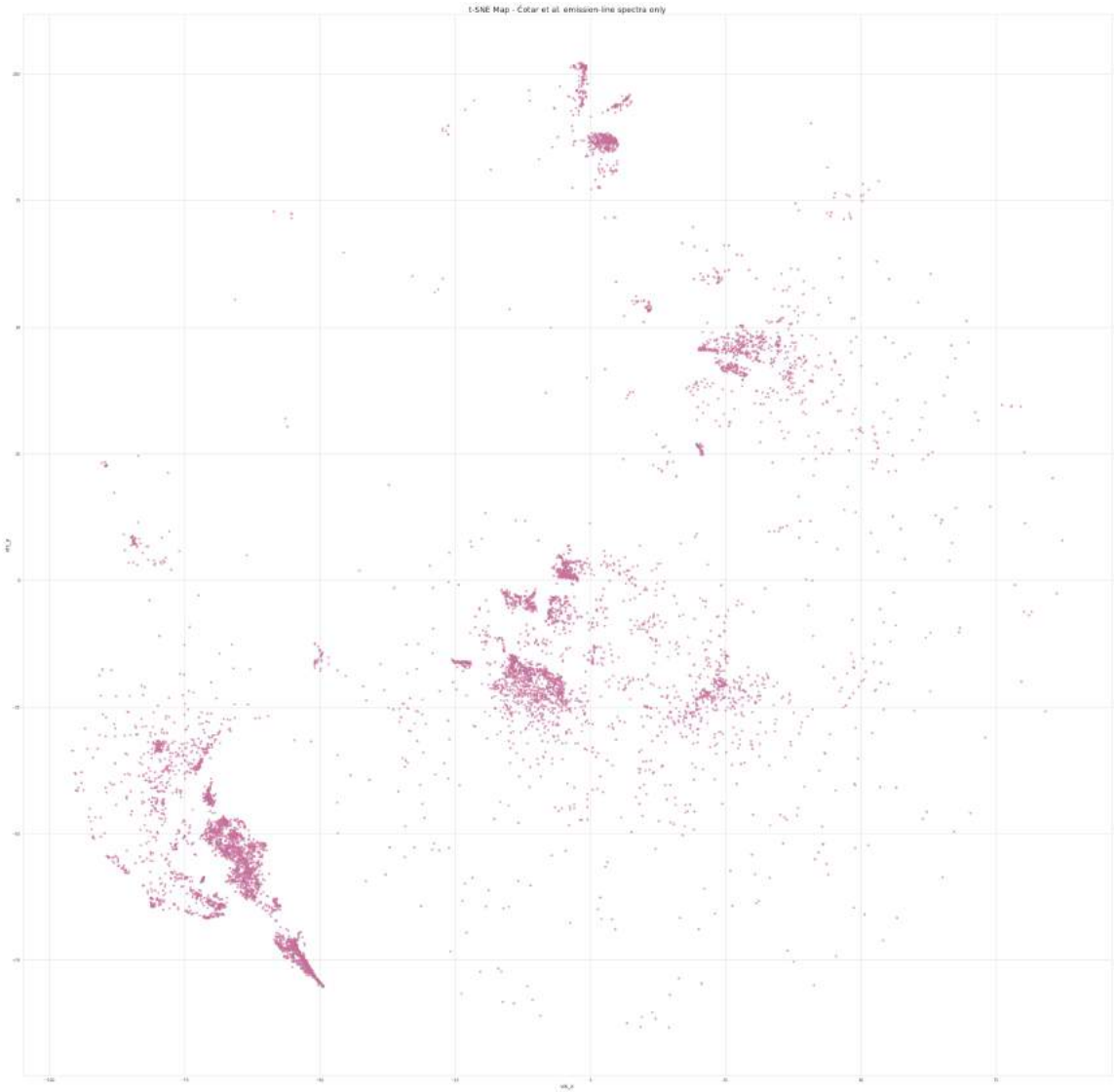

FIGURE A.2: The Hα emission-line spectra identified by Čotar et al. (2021), projected to the same t-SNE map as Figure A.1, but excluding the presumably "typical" spectra. Note that the emission-line spectra are distributed across numerous clumps, as well as scattered through a large part of the map.

A.2 Appendix - Chapter 6

Presented below are clusters of emission-line spectra identified by the method presented in Chapter 6. Over-classification is able to split double-peaked emission line spectra into two classes seen below in Figures A.3 and A.4. Figures A.5, A.6, A.7 and A.8 present additional results of the method discussed in Chapter 6.

Figure A.9 shows the equivalent width (EW) distribution of the catalogue released by Čotar et al., which was used as a point of comparison to the EW distribution for the emission-line spectra identified in Chapter 6.

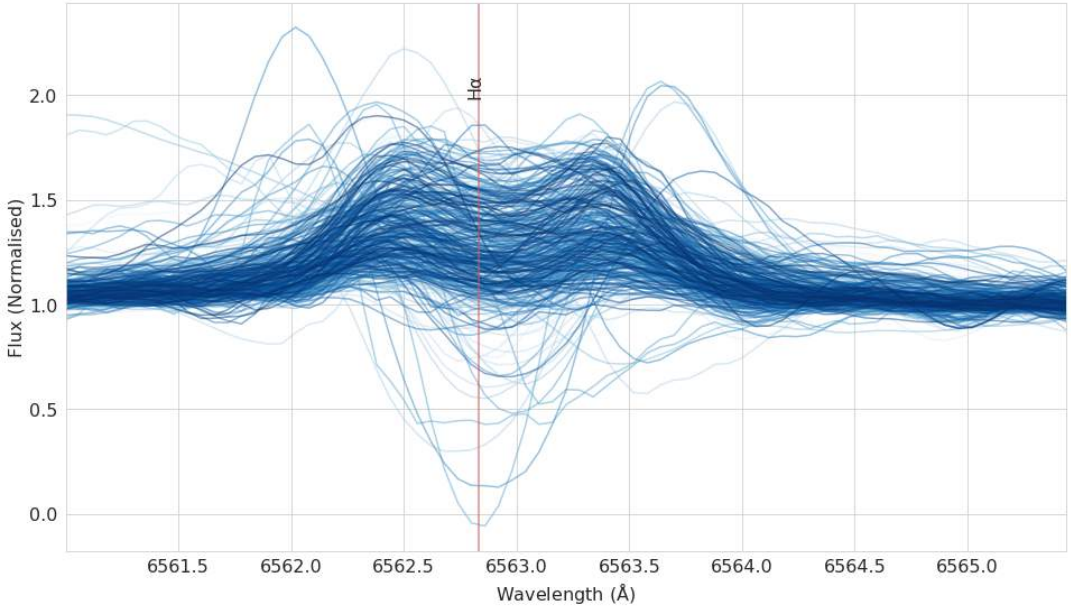

FIGURE A.3: A cluster of double-peaked spectra identified using DTW in the sample provided by Čotar et al. (2021).

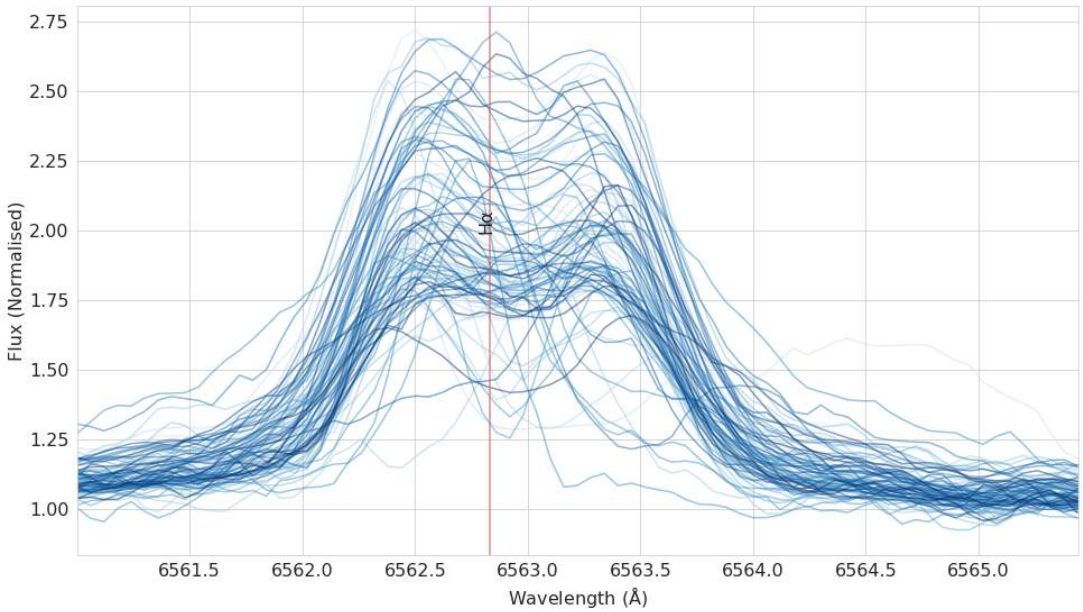

FIGURE A.4: A cluster of double-peaked spectra identified using DTW in the sample provided by Čotar et al. (2021).

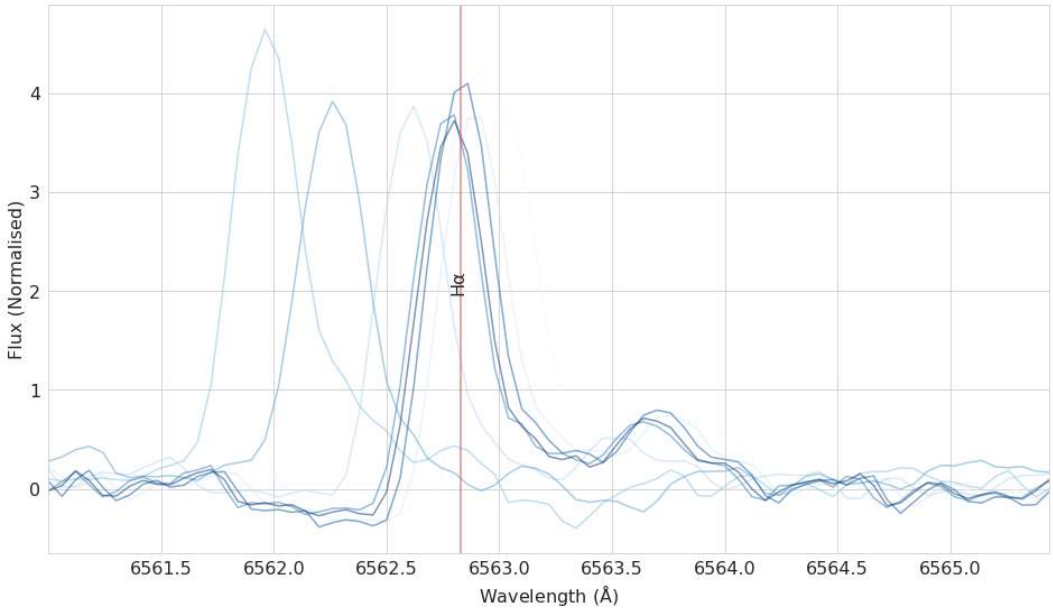

FIGURE A.5: Ensemble plot of 8 spectra with emission-lines superimposed on absorption

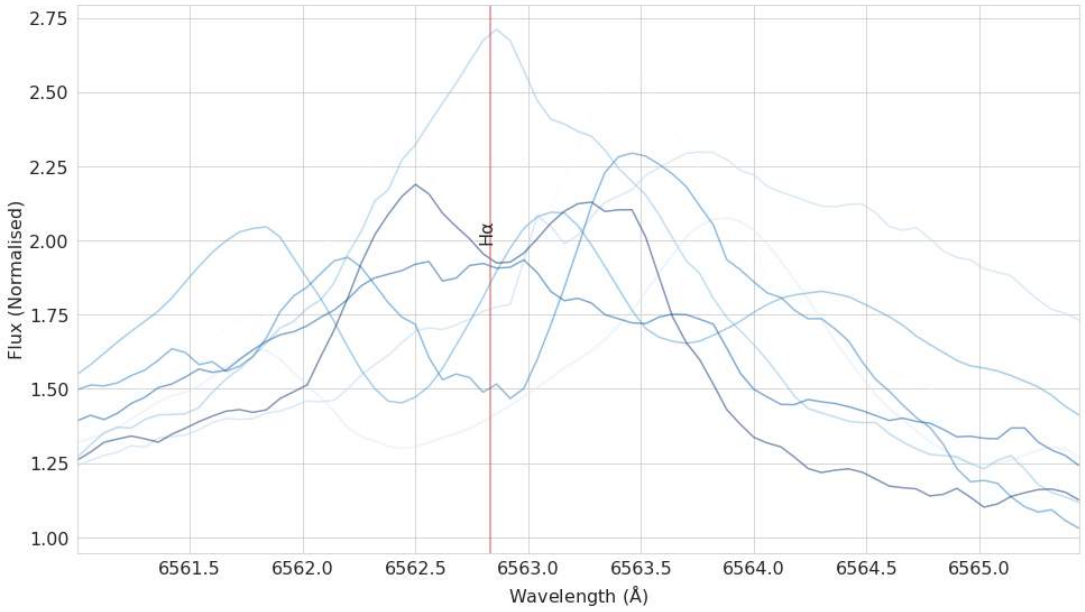

FIGURE A.6: Ensemble plot of 8 spectra with similar morphologies separated by DTW.

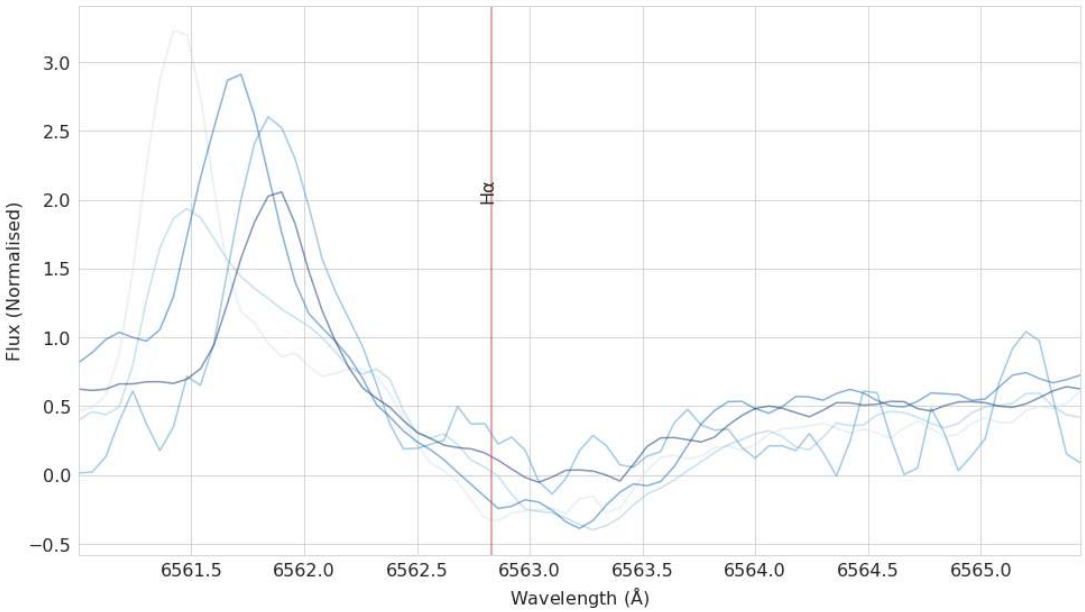

FIGURE A.7: Ensemble plot of 6 spectra with similar morphologies separated by DTW.

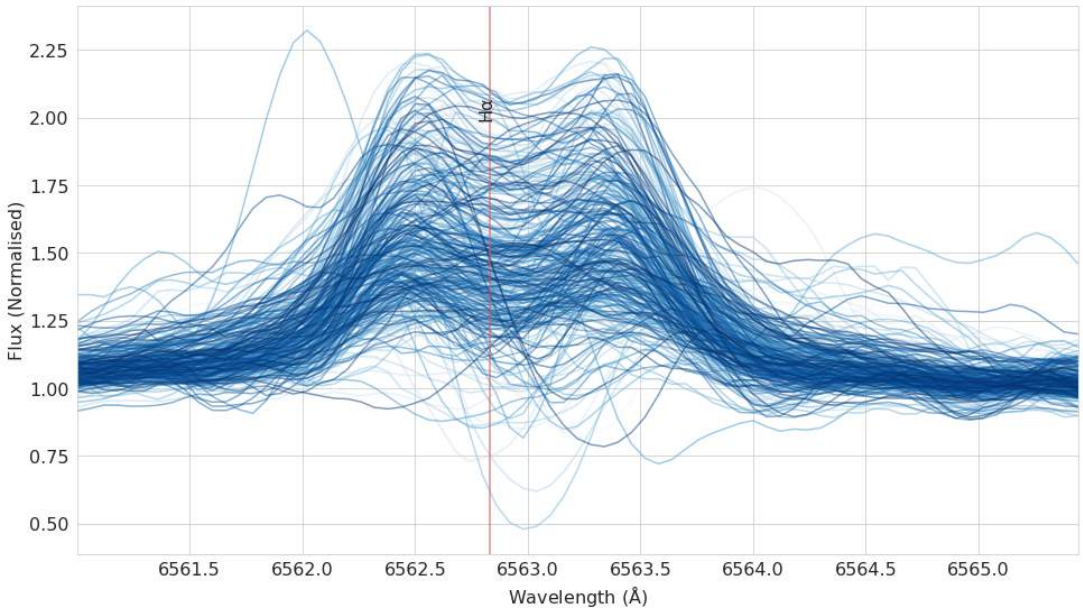

FIGURE A.8: Ensemble plot of 352 spectra with similar morphologies separated by DTW.

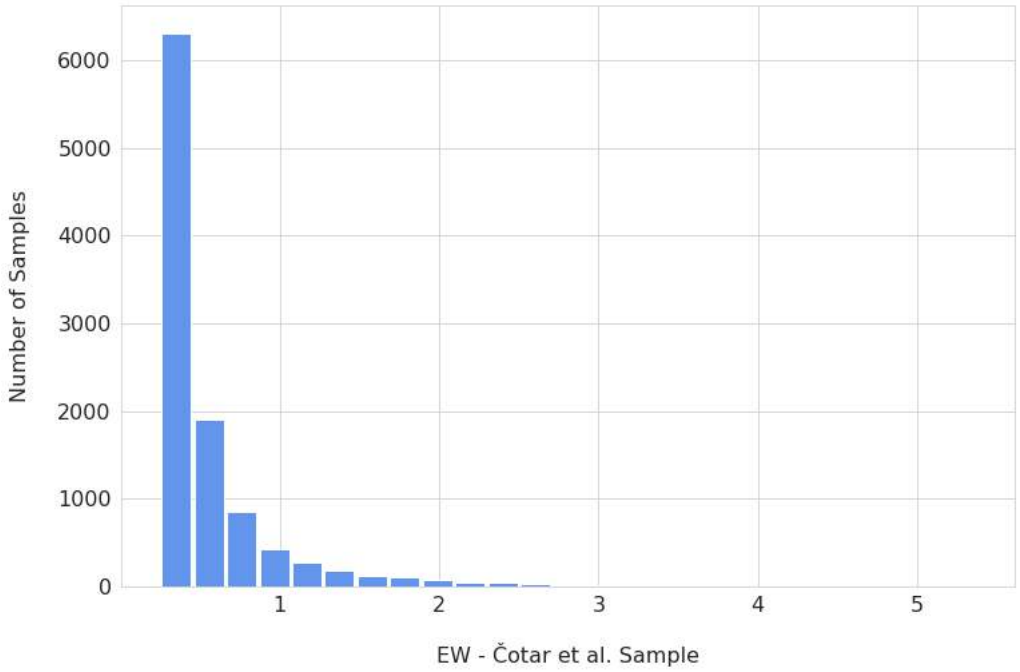

FIGURE A.9: The equivalent width (EW) distribution of the inverted difference spectra of the emission-line stars provided by Čotar et al. (2021). Here the spectra were selected such that EW>0.25. Note that this sample contains additional spectra not in GALAH DR3.

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/

Amarsi, A. M., Nordlander, T., Barklem, P., et al. 2018, Astronomy & Astrophysics, 615, A139

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, , 558, A33, doi: 10.1051/0004-6361/201322068

Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, , 156, 123, doi: 10.3847/1538-3881/aabc4f

Astropy-Specutils Development Team. 2019, Specutils: Spectroscopic analysis and reduction, Astrophysics Source Code Library, record ascl:1902.012. http://ascl.net/1902.012

Beals, C. S. 1953, Publications of the Dominion Astrophysical Observatory Victoria, 9, 1

Bergemann, M., Serenelli, A., Schönrich, R., et al. 2016, Astronomy & Astrophysics, 594, A120

Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, The Astronomical Journal, 154, 28

Bonito, R., Prisinzano, L., Guarcello, M., & Micela, G. 2013, Astronomy & Astrophysics, 556, A108

Buder, S., Sharma, S., Kos, J., et al. 2021, Monthly Notices of the Royal Astronomical Society, 506, 150

Campello, R. J., Moulavi, D., & Sander, J. 2013, in Pacific-Asia conference on knowledge discovery and data mining, Springer, 160–172

Carnall, A. 2017, arXiv preprint arXiv:1705.05165

Cayrel, R., Van't Veer-Menneret, C., Allard, N., & Stehlé, C. 2011, Astronomy & Astrophysics, 531, A83

Chollet, F., et al. 2015, Keras, GitHub. https://github.com/fchollet/keras

Čotar, K., Zwitter, T., Traven, G., et al. 2021, Monthly Notices of the Royal Astronomical Society, 500, 4849

De Groot, M., & Sterken, C. 2001, Publications of the Astronomical Society of the Pacific, 113, 258

De Silva, G. M., Freeman, K. C., Bland-Hawthorn, J., et al. 2015, Monthly Notices of the Royal Astronomical Society, 449, 2604

Eisenstein, D. J., Annis, J., Gunn, J. E., et al. 2001, The Astronomical Journal, 122, 2267

Elliott, A., Richardson, N. D., Pablo, H., et al. 2022, Monthly Notices of the Royal Astronomical Society, 509, 4246

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. 1996, 96, 226

Gaia-Collaboration, Brown, A. G. A., Vallenari, A., et al. 2021, Astronomy &amp Astrophysics, 650, C3, doi: 10.1051/0004-6361/202039657e

Garcia-Dias, R., Prieto, C. A., Almeida, J. S., & Ordovás-Pascual, I. 2018, Astronomy & Astrophysics, 612, A98

Giribaldi, R. E., Ubaldo-Melo, M. L., de Mello, G. P., et al. 2019, Astronomy & Astrophysics, 624, A10

Goodfellow, I., Bengio, Y., & Courville, A. 2017, Cambridge Massachusetts, 321

Hackeling, G. 2017, Mastering Machine Learning with scikit-learn (Packt Publishing Ltd)

Halevy, A., Norvig, P., & Pereira, F. 2009, IEEE intelligent systems, 24, 8

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. 2009, The elements of statistical learning: data mining, inference, and prediction, Vol. 2 (Springer)

He, K., Zhang, X., Ren, S., & Sun, J. 2015, in Proceedings of the IEEE international conference on computer vision, 1026–1034

Hogg, D. W., Bovy, J., & Lang, D. 2010, arXiv preprint arXiv:1008.4686

Hou, W., Luo, A.-L., Hu, J.-Y., et al. 2016, Research in Astronomy and Astrophysics, 16, 006

Hutchings, J. 1969, Monthly Notices of the Royal Astronomical Society, 144, 235

Itakura, F. 1975, IEEE Transactions on acoustics, speech, and signal processing, 23, 67

Kasai, E. K. 2013, Master's thesis, University of Cape Town

Kingma, D. P., & Ba, J. 2014, arXiv preprint arXiv:1412.6980

Kohoutek, L., & Wehmeyer, R. 1999, Astronomy and Astrophysics Supplement Series, 134, 255

Kos, J., Lin, J., Zwitter, T., et al. 2017, Monthly Notices of the Royal Astronomical Society, 464, 1259

Kruskal, J. B. 1983, SIAM review, 25, 201

Kuo, F. Y., & Sloan, I. H. 2005, Notices of the AMS, 52, 1320

MacQueen, J., et al. 1967in , Oakland, CA, USA, 281–297

Martell, S. L., Sharma, S., Buder, S., et al. 2017, , 465, 3203, doi: 10.1093/mnras/stw2835

Mizumoto, M., Kobayashi, N., Hamano, S., et al. 2018, Monthly Notices of the Royal Astronomical Society, 481, 793

Ness, M., Hogg, D. W., Rix, H.-W., et al. 2016, The Astrophysical Journal, 823, 114

Nielsen, A. 2019, Practical time series analysis: Prediction with statistics and machine learning (O'Reilly Media)

Reipurth, B., Pedrosa, A., & Lago, M. 1996, Astronomy and Astrophysics Supplement Series, 120, 229

Sakurada, M., & Yairi, T. 2014, in Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis, 4–11

Salvador, S., & Chan, P. 2007, Intelligent Data Analysis, 11, 561

Sharma, S., Stello, D., Buder, S., et al. 2018, Monthly Notices of the Royal Astronomical Society, 473, 2004

Sheinis, A., Barden, S., Birchall, M., et al. 2014, in Ground-based and Airborne Instrumentation for Astronomy V, Vol. 9147, SPIE, 344–365

Traven, G., Zwitter, T., Van Eck, S., et al. 2015, Astronomy & Astrophysics, 581, A52

Traven, G., Matijevič, G., Zwitter, T., et al. 2017, The Astrophysical Journal Supplement Series, 228, 24

Ulyanov, D. 2016, Multicore-TSNE, https://github.com/DmitryUlyanov/Multicore-TSNE, GitHub

Underhill, A. B. 1966, in The Early Type Stars (Springer), 213–225

Van der Maaten, L., & Hinton, G. 2008, Journal of machine learning research, 9

Van Winckel, H., Duerbeck, H., & Schwarz, H. 1993, Astronomy and Astrophysics Supplement Series, 102, 401

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nature Methods, 17, 261, doi: 10.1038/s41592-019-0686-2

Wittenmyer, R. A., Sharma, S., Stello, D., et al. 2018, The Astronomical Journal, 155, 84

Zhang, Y.-J., Hou, W., Luo, A.-L., et al. 2021, arXiv preprint arXiv:2107.00402

Zhao, G., Zhao, Y.-H., Chu, Y.-Q., Jing, Y.-P., & Deng, L.-C. 2012, Research in Astronomy and Astrophysics, 12, 723