Telluric Correction of M-dwarf Stars using Machine Learning

A dissertation submitted in
partial fulfillment for the degree of

Master of Technology

in

Computer Science

by

Sayak Rana
Roll no. - CS2427

under the supervision of

Dr. Sarbani Palit
Computer Vision and Pattern Recognition Unit (CVPRU)

Indian Statistical Institute, Kolkata

June, 2026

Certificate

This is to certify that the dissertation entitled **"Telluric Correction of M-Dwarf Stars Using Machine Learning"** submitted by **Sayak Rana** to the Indian Statistical Institute, Kolkata, in partial fulfillment of the requirements for the degree of Master Of Technology in Computer Science, is an authentic and genuine record of the research work carried out by the candidate under my supervision and guidance. I affirm that the dissertation has met all the necessary requirements in accordance with the regulations of this institute.

Sarbani Palit

Dr. Sarbani Palit 10.6.2026
CVPR Unit
Indian Statistical Institute
Kolkata - 700108
India

Declaration

I, **Sayak Rana**, holding Roll. No. **CS2427**, hereby declare that the content presented in the dissertation **Telluric Correction of M-Dwarf Stars using Machine Learning** represents the original work carried out by me for the degree of **Master of Technology in Computer Science** at **Indian Statistical Institute, Kolkata**. I affirm that no sections of this report have been sourced or copied from external references without proper attribution. I am aware that any instance of plagiarism or the use of unacknowledged materials from third parties will be treated with the utmost seriousness and consequences.

Sayak Rana
M.Tech. CS
Roll. No. CS2427
Indian Statistical Institute, Kolkata

Acknowledgments

I express my sincere appreciation to my supervisor, **Dr. Sarbani Palit**, Computer Vision and Pattern Recognition Unit (CVPRU) of Indian Statistical Institute, Kolkata for her invaluable guidance, encouragement, and unwavering support throughout this dissertation work. Her patience in answering all my doubts, providing insightful suggestions and constant motivation helped me throughout this journey. I am thankful for the trust and freedom she gave me to explore ideas independently and develop my understanding how research works.

I am deeply grateful to Ankita Sarkar, Senior Research Fellow at Indian Statistical Institute, for her generous assistance in collecting the required datasets, suggesting new brainstorming ideas, approaches and existing research papers related to this dissertation.

I would also like to extend my sincere gratitude to Dr. Arvind Singh Rajpurohit of the Physical Research Laboratory and Arindam Mal of the Indian Space Research Organisation (ISRO) for their generous support in resolving my physics-related doubts, clarifying fundamental concepts, and sharing their valuable domain knowledge.

I would also like to thank all the faculty members of the Indian Statistical Institute for their invaluable teaching and guidance. The concepts and knowledge gained through their courses provided a strong foundation for understanding many of the ideas and methodologies used in this dissertation.

I would also like to acknowledge the researchers, developers, and contributors whose publicly available datasets, models, software, and scientific resources formed an essential foundation for this work.

Lastly, I would like to express my heartfelt gratitude to my family for their endless support, encouragement, and understanding throughout the course of this work. I am equally grateful to my friends for their help, motivation, and constant encouragement. Their support played an important role in helping me successfully complete this dissertation.

Abstract

The study of M-dwarf stars is of prime scientific interest to us because of their closer habitable zones and the favorable conditions they offer for exoplanet detection. However, telluric contamination of the ground-based spectra results in sharp absorption lines, which makes their study cumbersome. Removing this contamination is necessary for estimating key stellar parameters.

The central contribution is a one-dimensional Convolutional Neural Network (CNN) that retrieves the four atmospheric parameters governing telluric absorption: pressure, temperature, humidity, and airmass. These predicted parameters are passed to Telfit which produces an estimated telluric spectrum. The observed spectrum is then divided by this estimated telluric spectrum to obtain the telluric corrected spectrum. As this network is trained exclusively on synthetic spectra, a domain gap exists at inference. Two domain-adaptation strategies are evaluated: a CycleGAN following the CycleStarNet framework for explicit synthetic-to-real translation and a Domain-Adversarial Autoencoder (DAAE) that learns domain-invariant spectral representations.

The Domain-Adversarial Autoencoder (DAAE) achieves the lowest loss against a telluric corrected CARMENES reference spectrum outperforming all other model variants. The CycleGAN fails due to discriminator collapse under severe class imbalance between the number of real and synthetic spectra.

Keywords: Telluric correction, M-Dwarf, CARMENES, CNN, CycleGAN, CycleStarNet, Domain-Adversarial Autoencoder (DAAE).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 M-Dwarf Stars

M-dwarfs are one of the most abundant types of star in the Galaxy. Their effective temperatures range from 2400 K to 3900 K, and radii range from 0.08 to 0.60 times the radius of the Sun. Thus, their habitable zones are closer compared to other stars because of their cooler temperatures and smaller radii. This makes them favorable for exoplanet detection.

This motivated the development of high-resolution spectrographs like CARMENES [1] - (Calar Alto High Resolution Search for M dwarfs with Exoearths with Near-infrared and optical Échelle Spectrographs) for M-dwarfs. It is installed at a 3.5 m telescope at Calar Alto Observatory in Spain and includes two regions: a visible region covering 5200 Å to 9600 Å with a spectral resolution of 94,600, and a near-infrared region spanning 9600 Å to 17100 Å at resolution 80,410. The two regions are equipped to provide high precision radial velocity measurements. The observed data that is used in this report have been taken from the CARMENES visible region (with spectral resolution of 94,600).

1.2 Telluric contamination

Before the light from a star reaches the ground-based telescope, it passes through the Earth's atmosphere. Now the Earth's atmosphere contains molecules such as oxygen (O_2), water vapor (H_2O), carbon dioxide (CO_2) and methane (CH_4) which absorb the incoming light at certain wavelengths. This absorption by these molecules leads to additional absorption lines at those particular wavelengths. These additional absorption lines in the ground-based observed spectra induced by the Earth's atmosphere are known as **"Telluric lines"**. These lines thus cause contamination on the observed

spectra which otherwise would have been a clean spectra. This phenomenon is known as telluric contamination.

Different atmospheric molecules absorb light at different wavelengths, thus producing telluric absorption bands with different depths and widths throughout the whole spectrum. The depth and width of the telluric lines depend heavily on the concentration of the atmospheric molecules. Higher concentrations result in deeper and broader absorption lines (telluric features). Among all molecules water vapor (H_2O) and oxygen (O_2) are the major contributors to telluric absorption. Also, the concentration of water vapor (H_2O) varies widely with Earth's weather conditions, location and time. Thus the contribution of water vapor (H_2O) to the telluric spectrum is really difficult to model and predict accurately.

The strength of the telluric absorption is related to four atmospheric parameters, such as temperature, pressure, relative humidity, and airmass (the length of the path that the incoming starlight would travel through the atmosphere). Airmass increases as the telescope points farther away from the zenith (i.e as the zenith angle increases) which again causes the starlight to pass through a larger distance in the Earth's atmosphere. Changes in any of these parameters may affect the depth, width of telluric absorption lines observed in the spectrum.

Telluric contamination hinders the study of M-dwarfs. M-dwarfs are cool and their photospheres contain molecular species like TiO, VO, and H_2O which result in broad absorption bands in the red and near-infrared regions of a spectrum. Now these stellar molecular bands fall almost at the same wavelengths as the telluric bands, specially the telluric lines due to water vapor (H_2O) and oxygen (O_2). Thus the stellar and telluric lines in these regions are so closely intertwined in these regions for M-dwarf stars that separating them becomes a really challenging task.

If the telluric lines are not eliminated, they will destroy all the measurements extracted from the spectrum. All the primary stellar parameters of M-dwarfs, such as effective temperature $T_{\rm eff}$, surface gravity $\log g$, and metallicity [M/H] are actually estimated from the strengths,depths and widths of the spectral absorption lines. Telluric lines (features) mixed with the stellar lines (features) contaminate the spectrum and thus introduce bias in the derived stellar parameters. Inaccurate estimation of primary stellar parameters leads to wrong estimation of other secondary parameters that depend on them, such as stellar age, mass, radius and magnetic activity. Telluric contamination also leads to wrong estimation of radial velocity. Even today, the usual

approach is to mask out those wavelength ranges entirely and discard the spectral regions which involves heavy telluric contamination. This eliminates the bias and inaccurate stellar parameter estimation but on the other hand we discard a significant fraction of the available stellar information which could have been useful. Thus, removing the telluric contamination entirely, instead of masking it is necessary for making full use of the observed spectrum.

1.3 Physical Basis of Telluric contamination

Each observed spectrum by the ground-based telescope contains the contribution from the star (stellar feature) as well as from the Earth's atmosphere (Telluric feature). This observed spectrum could be written as [2],

$$F_{\text{obs}}(\lambda) = \left[F_{\star}(\lambda) \cdot T(\lambda)\right] * I(\lambda) \cdot Q(\lambda), \tag{1.1}$$

where $F_{\star}(\lambda)$ denotes the intrinsic stellar spectrum, $T(\lambda)$ denotes telluric transmittance, $I(\lambda)$ denotes the instrumental line-spread function, $Q(\lambda)$ denotes the instrumental throughput and $*$ sign represents convolution operation.

The instrumental effects are ignored in this work that is we assume an ideal spectrograph where the throughput is uniform, ($I(\lambda) \to \delta(\lambda)$, $Q(\lambda) \to 1$). Thus , Equation (1.1) reduces to,

$$F_{\text{obs}}(\lambda) = F_{\star}(\lambda) \cdot T(\lambda), \tag{1.2}$$

where $T(\lambda) \in [0, 1]$ is the telluric transmittance.

Furthermore, the telluric transmittance itself is determined by the combined absorption of multiple molecular species,

$$T(\lambda) = \prod_{i=1}^{K} t_i(\lambda), \tag{1.3}$$

where $t_i(\lambda)$ is the transmission spectrum of the i-th atmospheric constituent. These include water vapor (H_2O), oxygen (O_2), carbon dioxide (CO_2), methane (CH_4) etc.

The main goal of the telluric correction is to obtain the intrinsic stellar spectrum

by estimating and eliminating the telluric contamination:

$$\hat{F}_{\star}(\lambda) = \frac{F_{\text{obs}}(\lambda)}{\hat{T}(\lambda)}. \tag{1.4}$$

Here $\hat{F}_{\star}(\lambda)$ denotes the estimated stellar spectrum and $\hat{T}(\lambda)$ denotes the estimated telluric spectrum.

This is a difficult task, as the best fit $\hat{T}(\lambda)$ is not known beforehand. In standard solar-type stars, there exists sub-regions where intrinsic stellar spectrum is smooth without strong stellar absorption lines. In these sub-regions, the entire contribution to observed spectrum comes from the telluric spectrum. Thus for these stars $T(\lambda)$ can be estimated directly from the observation. But in M-dwarfs as both the telluric and stellar lines lie in the same region, distinguishing between them is a difficult task and thus extracting $\hat{T}(\lambda)$ is cumbersome.

1.4 Overview of the Report

The remainder of this report is organized in the following ways: Chapter 2 describes the observational and synthetic datasets used in the proposed work. Chapter 3 reviews existing approaches to telluric correction. Chapter 4 describes the full methodology of my proposed work in detail. Chapter 5 presents the experimental results of my proposed pipeline. Chapter 6 closes with a summary, includes the challenges faced and directions for future work.

Chapter 2

Datasets

2.1 CARMENES observed spectra

This work performs telluric correction on M-dwarf spectra from the CARMENES instrument [1]. The visible region has a resolving power of $R \approx 94{,}600$ and it ranges from 520–960 nm.

The 2018 CARMENES dataset serves as the primary input. It comprises 382 M-dwarf stars but only 41 were considered after preprocessing. The spectra used here are telluric contaminated spectra (that is they carry both stellar and telluric features). The proposed work after preprocessing operates on two spectral chunks of these spectra: **Chunk 1** (7550–7750 Å, O_2 band) and **Chunk 2** (8300–8500 Å, water vapor band).

The 2023 CARMENES dataset serves as the telluric corrected reference at the time of evaluation. Same 41 telluric corrected stars were considered.

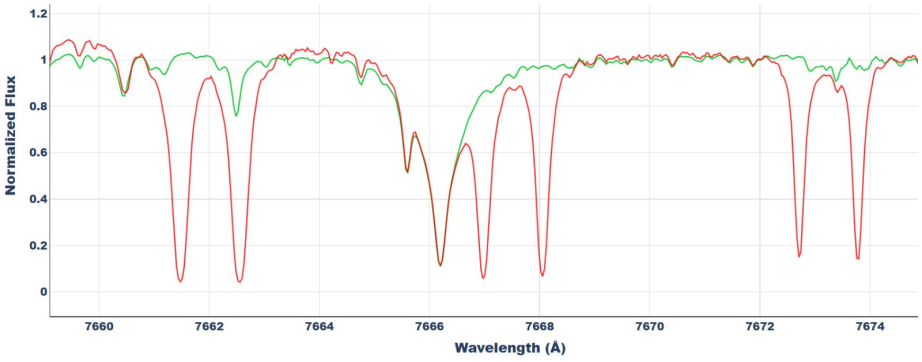

Figure 2.1: 2018 observed telluric contaminated spectrum (red; stellar + telluric) versus the telluric corrected reference spectrum (green; stellar).

2.2 BT-Settl synthetic stellar library

BT-Settl [5] provides synthetic stellar spectra. Each spectrum consists of three stellar parameters effective temperature ($T_{\rm eff}$), surface gravity ($\log g$) and metallicity ([M/H]) as the labels. A total of 214 spectra were available. A neural emulator was later trained to generate synthetic spectra for any permutation-combination of these parameters using this dataset.

2.3 Synthetic telluric grid from Telfit

Telfit [4] Synthetic Spectra Generator was used to produce a synthetic telluric spectrum for a given atmospheric state (P, T, H, A). Later each spectrum was degraded to CARMENES VIS resolution of $R = 94{,}600$ and the two spectral chunks were extracted.

This synthetic grid was built in two stages. For that the minimum and maximum values of all four atmospheric parameters were extracted from the 41 observed FITS headers. A uniform grid of 10 values per parameter was constructed within these bounds, yielding $10^4 = 10{,}000$ spectra (Table 2.1).

In the second stage, an additional 10,000 spectra were generated to cover the parameter space outside the initial range. These were sampled using Latin Hypercube Sampling (LHS) over a wider parameter range (Table 2.2) excluding points that were already covered at the first stage. The two sets were merged to form a final training dataset of 20,000 synthetic telluric spectra on which all models were later trained. 80/20 split was done on the dataset.

Table 2.1: Atmospheric parameter ranges for the 10,000 synthetic telluric grid. Maximum and Minimum values were obtained from the 41 observed CARMENES FITS headers.

Parameter	**Minimum**	**Maximum**
Pressure (hPa)	767.0	787.1
Temperature (K)	270.85	294.55
Relative Humidity (%)	13.0	92.0
Airmass	1.0003	2.1665

Table 2.2: Atmospheric parameter ranges for the additional 10,000 LHS-sampled spectra. Combined with the first stage, this forms the full 20,000 training set.

Parameter	**Minimum**	**Maximum**
Pressure (hPa)	741.0	800.0
Temperature (K)	265.0	305.0
Relative Humidity (%)	1.0	99.0
Airmass	1.0	3.0

Chapter 3

Related Works

Three widespread methods are used for telluric correction.

First, the classical method relies on obtaining a standard hot star at similar airmass as the observed spectrum. Then the telluric transmission feature is estimated by dividing the observed spectrum with this spectrum [2]. Although it looks quite straightforward but this method requires more telescope time and it also leaves errors from the standard star's own stellar features.

The second method generates the best fit synthetic telluric spectrum using radiative-transfer code such as Molecfit [3] and Telfit [4] which fits with the observed spectrum. This is the current state of the art physics based method still used. It can achieve high accuracy but it depends heavily on an accurate knowledge of all the atmospheric conditions at the time of observation which usually remains unavailable.

The third class of telluric correction methods uses a data-driven approach, which relies on machine learning to understand the relationship between observed and telluric contaminated spectra across large training sets [7, 8].

The telluric correction pipeline developed in this dissertation belongs to the third category. It combines a Convolutional Neural Network (CNN) based framework for atmospheric parameter estimation with domain-adaptation techniques to reduce the gap between synthetic training data and real observed data.

Chapter 4

Methodology

4.1 Overview

The pipeline is divided into four sequential stages, illustrated in Figure 4.1.

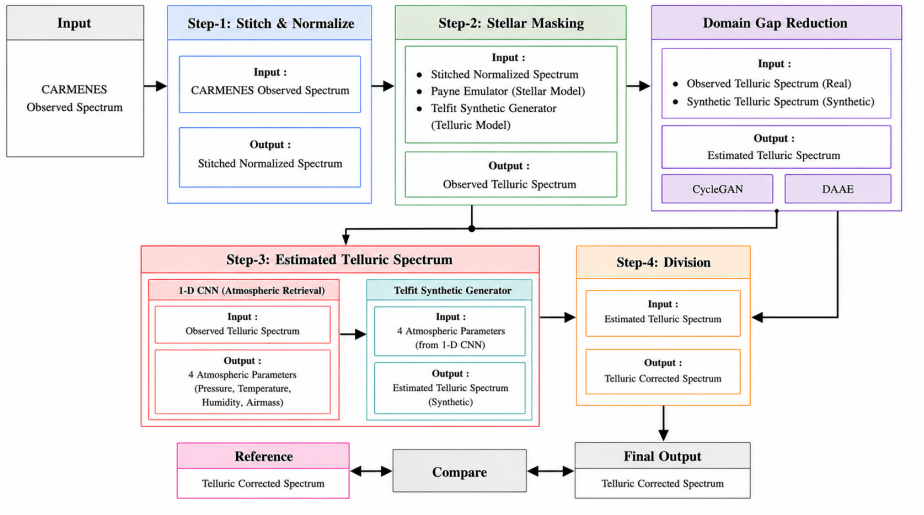

Figure 4.1: Overview of the telluric correction pipeline.

As depicted in the figure, in **Step 1**, for each raw CARMENES VIS observation we stitch the individual echelle orders to a single continuous spectrum. After that we normalize it to remove the instrumental continuum. The normalized spectrum is then passed to Step 2 for further preprocessing.

In **Step 2**, BERV and RV corrections are applied to bring the stitched normalized spectrum into the stellar rest frame. After that BT-Settl synthetic grid is used to train the *Payne* stellar emulator. The *Payne* emulator thus produces the pure synthetic stellar spectra. Next as depicted in the figure the *Payne* stellar emulator and the Telfit synthetic telluric model are both used jointly to identify and suppress stellar absorption features. After that reverse BERV and RV corrections are applied to bring

the stellar masked spectra to observed frame. This results in masking of the stellar regions in observed spectrum to give us an observed telluric spectrum.

In **Step 3**, the output from Step 2 is directly fed into the trained 1-D CNN and it predicts the atmospheric parameters. These are then passed back to Telfit synthetic spectra generator to generate a physically consistent estimated telluric spectrum.

However, there exists a domain gap between Step 2 output that is observed telluric spectra (derived from real observation) and Telfit generated synthetic telluric spectra. Two strategies are explored to bridge this *real-to-synthetic domain gap* - first using a **CycleGAN**, which learns a translation between the real and synthetic spectral domains, and later a **DAAE**, which trains the encoder to produce domain-invariant features, now the estimated synthetic telluric spectrum produced from both of these can directly be fed as input to 1-D CNN and then the same process continues or this can directly be considered for Step 4 - Division.

Finally, **Step 4** divides the stitched-normalized observation by this estimated telluric spectrum to yield the telluric corrected stellar spectrum, which is then compared against the telluric corrected spectrum (Reference) of the same target star.

4.2 Step 1 - Stitch & Normalize

4.2.1 Dataset and Observation Matching

The **VIS regions** of the CARMENES spectrograph cover a wavelength range 5200–9600 Å with a resolving power of $R = 94{,}600$. Each raw observation is stored as a two-dimensional array of 61 echelle orders with up to 3699 flux values per order, accompanied by a signal-to-noise (SNR) array for each wavelength point.

The observations are drawn from the 2018 CARMENES survey consisting of 382 M-dwarf stars. However , the corresponding reference telluric corrected spectrum is not available for all 382 M-dwarf stars. So, to ensure a fair and controlled evaluation, each 2018 observed spectrum is paired with a 2023 observation of the *same* star (from CARMENES 2023 dataset) obtained on the *same* night with the same date, airmass and instrumental setup from their respective FITS header. This cross-matching initially yielded **62 observation pairs**, which form the input to Step 1. However, during Step 1, some matched observations were found to be inappropriate after applying the BERV and RV corrections and then aligning them with the BT-Settl model grid.

These observations were later discarded. As a result, the final working dataset consists of **41 observation pairs**, which are used in all later stages of the pipeline. Of these 41 pairs, all 41 are used for domain adaptation training in the CycleGAN experiment. For the DAAE, 34 are used for training and the remaining 7 are held out as unseen test stars for generalization evaluation.

4.2.2 Order Stitching

Each raw VIS observed spectrum consists of 61 overlapping echelle orders. Order Stitching ensures a single continuous 1-D spectrum in the output. In the overlapping regions, the flux from the new order is first interpolated onto the wavelength grid of the existing spectrum, and the two are then combined as a weighted average; giving more weight to the order with the higher SNR at each wavelength point. The non-overlapping portion of each order is appended directly. This results in a single continuous 1-D observed spectrum.

4.2.3 Continuum normalization

The stitched spectrum carries varying continuum envelope from the instrumental blaze function. This is removed by repeating following three step cycle three times.

1. **Smoothing and normalization:** Each spectrum is smoothed using a Savitzky-Golay filter. It reduces small-scale noise but it preserves the underlying shape of spectral features. It is preferred over a simple boxcar as it avoids edge effects retaining narrow absorption structures. The window size is halved at each successive iteration so that it is able to capture finer details of the continuum.

2. **Continuum Subtraction:** After this, the smoothed estimate is subtracted from the spectrum. This helps to isolate the absorption features and eliminate the broad global trends in the flux.

3. **Polynomial Approximation:** Then, a linear polynomial is fitted over the shoulders, on either side of each detected absorption feature. This helps to estimate the local continuum level. This fitted baseline finally replaces the absorption region which prevents deep lines from biasing the continuum estimate in subsequent iterations.

Repeating this three step cycle three times yields a robust, smooth continuum model. The normalized flux is finally obtained by dividing the original stitched spectrum by this continuum model, thus the line-free regions settle near unity.

Thus the final output of Step 1 is a continuum-normalized, stitched 1-D spectrum for each of the 41 observations. It has both stellar and telluric absorption features. Lastly, this stitched normalized spectrum serves as the input to all subsequent steps of the pipeline.

4.3 Step 2 - Stellar Masking and obtaining observed telluric spectra

4.3.1 Preprocessing for Step 2

BERV and RV Correction

The stitched-normalized 2018 CARMENES observation is in the observer (topocentric) frame, while the BT-Settl synthetic stellar spectra are computed in the stellar rest frame. Now, to make the stellar mask work properly, the observed stellar absorption dips must align perfectly with the synthetic BT-Settl model. Therefore two Doppler corrections (BERV and RV) are applied.

The effect of the Earth's orbital motion around the Solar System barycentre must be removed to place the observed spectrum in a common reference frame. This Doppler shift in the observed wavelengths is corrected using the multiplicative factor,

$$f_{\mathrm{berv}} = 1 + \frac{v_{\mathrm{berv}}}{c}, \tag{4.1}$$

where v_{berv} is obtained from the FITS header (`HIERARCH CARACAL BERV`) and c is the speed of light.

After correcting the Earth's motion, the spectrum is further shifted into the stellar rest frame by accounting for the star's systemic motion along the line of sight. This Doppler correction is performed using the **stellar radial velocity (RV)** obtained from the matched 2023 VIS FITS header using the equation:

$$f_{\mathrm{rv}} = \sqrt{\frac{1 - v_{\mathrm{rv}}/c}{1 + v_{\mathrm{rv}}/c}}, \tag{4.2}$$

where v_{rv} is the stellar radial velocity obtained from the matched 2023 FITS header. The corrected wavelength grid is obtained by multiplying the observed wavelengths by both berv and rv factors, as follows:

$$\lambda_{\mathrm{corr}} = \lambda_{\mathrm{obs}} \times f_{\mathrm{berv}} \times f_{\mathrm{rv}}. \tag{4.3}$$

Without these corrections, the stellar absorption lines in the observed spectrum would not align with those in the synthetic model, preventing accurate stellar masking in further steps.

Spectral Chunking and Grid Standardization

Although the stitched-normalized spectra cover the full VIS wavelength range, but the subsequent analysis in further steps focuses only on two wavelength regions that are strongly affected by telluric absorption:

- **Chunk 1**: 755–775 nm
- **Chunk 2**: 830–850 nm

These two chunks were selected because they contain some of the strongest telluric features in the CARMENES VIS spectrum. Thus it provides a challenging test bed for evaluating the performance of my proposed telluric correction pipeline. Furthermore, restricting the analysis to these heavily contaminated segments allows the effectiveness of the proposed method to be assessed under demanding conditions before extending it to the entire VIS spectrum range.

All observed spectra are interpolated onto fixed wavelength grids covering regions: 755–775 nm and 830–850 nm respectively with a step size of 0.0035 nm (matches the native wavelength grid of the CARMENES VIS spectra). Linear interpolation is used to map each observed spectrum onto these grids. Points outside the available wavelength range are filled with a normalized flux value of unity. From this stage onwards, all observed and synthetic spectra are represented on these common wavelength grids.This fixed grid is preserved throughout the pipeline.

BT-Settl Synthetic Spectra Preprocessing

BT-Settl model grid gives us a synthetic stellar spectra at a very high spectral resolution. Prior to the comparison of stellar dips of BT-Settl generated synthetic

spectrum and CARMENES observed spectrum some preprocessing steps are applied to the BT-Settl spectrum. They are :

1. **Spectral resolution degradation:** As mentioned BT-Settl spectra are at a high spectral resolution, far exceeding the CARMENES ($R \approx 94{,}600$). So, spectral degradation of each spectrum is performed. For that each spectrum is convolved with a wavelength dependent Gaussian kernel.
2. **Interpolation and normalization:** After this the degraded spectrum is re-sampled to the two CARMENES uniform chunks with a step size of (0.0035 nm). Then it is continuum normalized using the same iterative Savitzky–Golay procedure , described in Section 4.2.3.

Matching Each Star to the BT-Settl Grid

Each of the 41 observed stars has published stellar parameters $(T_{\rm eff}, \log g, [\mathrm{M/H}])$ compiled from the literature. As the BT-Settl library is a discrete grid, there is in general no model at exactly the right parameter combination for the observed star. The closest available grid spectrum for each star is thus computed by minimizing a weighted Euclidean distance:

$$d = \sqrt{\left(\frac{T_{\rm eff} - T_{\rm eff}^{\rm grid}}{100}\right)^2 + \left(\frac{\log g - \log g^{\rm grid}}{0.5}\right)^2 + \left(\frac{[\mathrm{M/H}] - [\mathrm{M/H}]^{\rm grid}}{0.5}\right)^2}, \qquad (4.4)$$

where $T_{\rm eff}^{\rm grid}$, $\log g^{\rm grid}$, and $[\mathrm{M/H}]^{\rm grid}$ denote the stellar parameters of a candidate BT-Settl model. The scaling factors in the denominator correspond to the typical spacing of the BT-Settl parameter grid. It normalizes the contribution of each parameter. The synthetic spectrum associated with the minimum distance d is selected as the reference stellar spectrum for the corresponding observed star.

4.3.2 The Payne stellar emulator

The BT-Settl grid consisted of 214 synthetic spectra sampled at coarse, discrete values of $(T_{\rm eff}, \log g, [\mathrm{M/H}])$. But a direct nearest-grid-point match produces a poor approximation of the true stellar spectrum, as the real stellar labels take continuous values. To address this specific problem, a *Payne* emulator [6] is trained. It is a neural network that takes any continuous label and predicts the corresponding flux vector.

The 214 files were first split into 171 training files and 43 held-out test files (80/20 split). Initially the *Payne* emulator was trained directly on 171 original BT-Settl spectra. Later to get better results, the training set was subsequently expanded via linear interpolation. By treating the 171 training spectra as anchor points new synthetic spectra at intermediate label values were synthesised. The flux at each wavelength point is a distance weighted average of the surrounding anchors. Interpolation was applied on a finer grid (25 K steps in $T_{\rm eff}$, 0.1 steps in both $\log g$ and [M/H]). Also, any point lying outside the parameter range of the 171 anchors was discarded. The emulator was re-trained on this expanded set and the quantitative improvement over the 214 file baseline is reported in Section 5.1.

The emulator architecture is taken from [6]. The three stellar labels ($T_{\rm eff}$, $\log g$, [M/H]) are scaled to [0, 1] and fed as input into two fully connected hidden layers of 300 neurons each with Leaky ReLU activation functions. The output layer is a single linear layer that predicts the normalized flux at every point on the wavelength grid. The model architecture is available in Figure 4.2.

The network is trained using the Adam optimizer with an L1 (mean absolute error) loss, a learning rate of 10^{-3}, and ℓ_2 weight decay of 10^{-5}. Five-fold cross-validation helps to determine the optimal stopping epoch. The model is then re-trained on the full training set for those many epochs with early stopping.

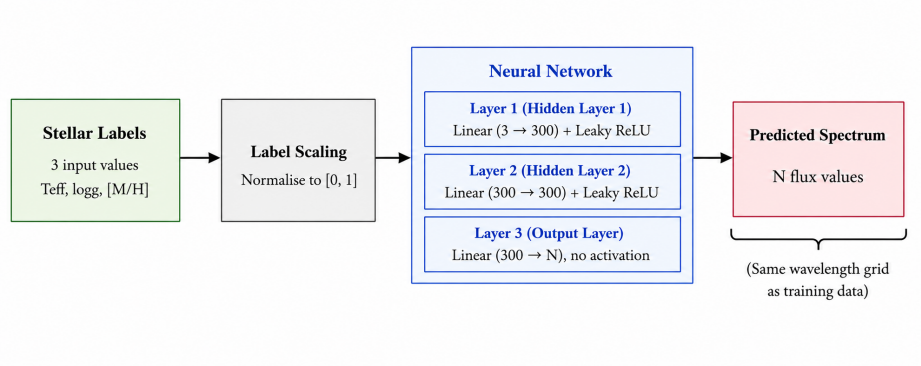

Figure 4.2: The *Payne* emulator. Scaled stellar labels ($T_{\rm eff}$, $\log g$, [M/H]) are fed into two hidden layers of 300 neurons with LeakyReLU activations. Finally, the linear output layer predicts the normalized flux at every wavelength point.

During inference, the *Payne* emulator is called with the known stellar parameters of each target star. It generates an appropriate stellar spectrum corresponding to the specific input combinations.

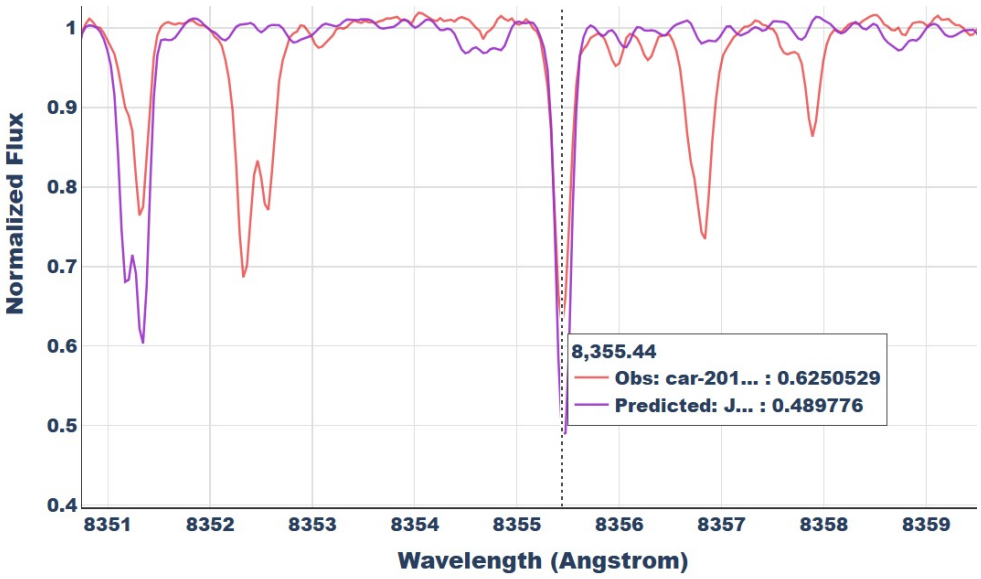

Figure 4.3: Observed CARMENES spectrum (red) overlaid with the *Payne*-predicted synthetic stellar spectrum (purple) in the 8350–8359 Å region. The synthetic spectrum contains only stellar absorption features while the observed spectrum carries both stellar and telluric absorption. Dips present in the synthetic but absent in the observed (e.g. near 8355 Å) indicate pure stellar features that are targeted for masking.

Three masking approaches were explored:

Approach 1: Threshold flattening: For each observed spectrum, the corresponding *Payne* emulator generated synthetic spectrum was used to locate stellar absorption regions. Wavelengths where the synthetic flux fell below a fixed threshold (0.90) was marked as stellar and the observed flux values in that region was replaced with 1.0. To ensure the complete coverage of each stellar feature, the mask was expanded by applying a small padding width on either side of the identified absorption region. But this process resulted in abrupt discontinuity and also this method failed to mask the entire region.

Approach 2: Division and spike suppression: The observed spectrum was at first divided by the *Payne* emulator generated synthetic spectrum. Ideally, this operation should remove the stellar absorption features. But, imperfect alignment (in depth and width) between stellar features of observed and synthetic resulted in spikes or dips. So the flux values in these regions were replaced with $\min(f_L, f_R)$, where f_L

and f_R denote the mean flux values at the left and right shoulders of the region. This produced noticeably cleaner results than threshold flattening.

Approach 3: Manual mask with interpolated filling (adopted): Each stellar region was manually identified by overlaying the *Payne*-predicted stellar spectrum on the observed spectrum for each star. Wherever there occurred a dip in both the synthetic and observed spectra, that region was marked as stellar and flagged for masking. The flux values in that region was then replaced by:

$$F_{\text{masked}}(\lambda) \;=\; f_L + (f_R - f_L)\,\frac{\lambda - \lambda_L}{\lambda_R - \lambda_L}, \qquad \lambda \in [\lambda_L,\, \lambda_R]. \tag{4.5}$$

where $F_{\text{masked}}(\lambda)$ is the interpolated flux assigned to the masked wavelength λ, while λ_L and λ_R denote the left and right boundaries of the stellar absorption region, respectively. The quantities f_L and f_R are the flux values of left and right region respectively.

Each masked region was cross-checked against a Telfit-generated synthetic telluric spectrum. If a significant absorption dip was present in the Telfit spectrum at the same wavelength the region was denoted telluric and was not considered for masking. This approach was the most reliable of the three and therefore it is adopted for all subsequent steps. Following masking, reverse BERV and RV corrections were applied to the wavelength axis to shift the masked spectrum back to the topocentric frame, aligning it with the Telfit-generated synthetic telluric spectra. Figure 4.4 shows the masked observation against several Telfit spectra.

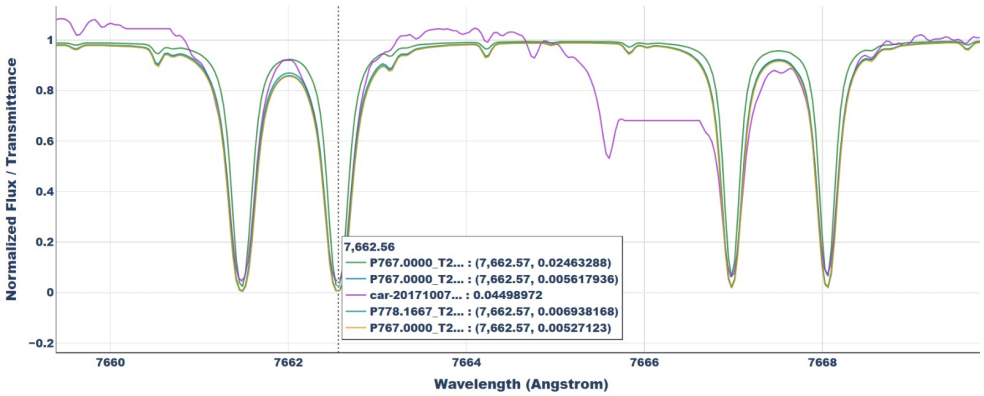

Figure 4.4: Masked observation (purple) against Telfit generated synthetic telluric spectra. Stellar dips are suppressed due to masking while telluric features remain preserved.

4.4 Step 3 - Generating estimated telluric spectrum

4.4.1 Objective

The observed telluric spectrum produced in Step 2 is given as input to Step 3. The main goal of Step 3 is to estimate the four atmospheric parameters - pressure (P), temperature (T), humidity (H) and airmass (A). These parameters are then fed into Telfit to generate a physically consistent and refined telluric transmittance spectrum.

4.4.2 1D-CNN

A one-dimensional convolutional neural network (1D-CNN) is employed to estimate the four atmospheric parameters directly from the observed telluric spectrum. The trained network predicts all four parameters from a single input spectrum.

Convolutional neural networks are well suited for this task as the telluric features are actually localized absorption features and the atmospheric parameters are responsible for producing the depth and width intensity of these features. Before training, the four atmospheric parameters (P, T, H, A) were transformed to a common numerical range. This prevents parameters with larger numerical values from dominating the optimization process. Then it is passed to the encoder:

Encoder: The encoder consists of four convolutional blocks.

The input is a single-channel of telluric spectrum and of length L, where L is the number of wavelength points. Thus the input is of shape (1, L). Each of the first three blocks applies a convolution followed by batch normalization, ReLU activation, and a max-pooling that halves the sequence length, which shrinks the spectrum length but extracts increasingly abstract spectral features.

The number of feature maps across the first three blocks grows from $32 \rightarrow 64 \rightarrow 128$, with convolutional kernel sizes of 15, 9, and 5 respectively. The larger kernels in the early blocks capture broader spectral features while smaller kernels in the later blocks refine the local structure.

Each max pooling layer uses a stride of 2, which is why the sequence length halves after each block. Thus the output of the first block is of shape (32, $L/2$), from the second is (64, $L/4$) and from the third is (128, $L/8$).

The fourth block applies a convolution that reduces the number of channels from 128 to 64 with batch normalization and ReLU, but no pooling is applied. So the sequence length is preserved and the output shape is (64, $L/8$).

Finally, an adaptive average pooling layer compresses each of the 64 channels to a fixed length of 64. This reduces the output to a fixed size feature vector of shape (64, 64). This is later flattened to 4,096 and passed to the regression head. Adaptive average pooling ensures output shape remains independent of the input length L. Thus even if the spectrum length changes the downstream linear layers requires no further modification.

Regression head: The regression head consists of three fully connected layers.

The encoder output is passed to the first fully connected layer which reduces it to 256 units. This is followed by ReLU activation function and a dropout of 0.3.

The second fully connected layer similarly reduces the 256 units to 64 units, followed by ReLU activation and dropout of 0.2.

The final layer maps the 64 units to 4 output values, one for each atmospheric parameter (P, T, H, A).

Lastly a sigmoid activation is applied that restricts all four outputs within (0, 1). This matches the (0, 1) range scaling applied before training.

Furthermore, this also ensures the predicted outputs to be within the training input range and not predict absurd values during inference. The architecture is illustrated in Figure 4.5.

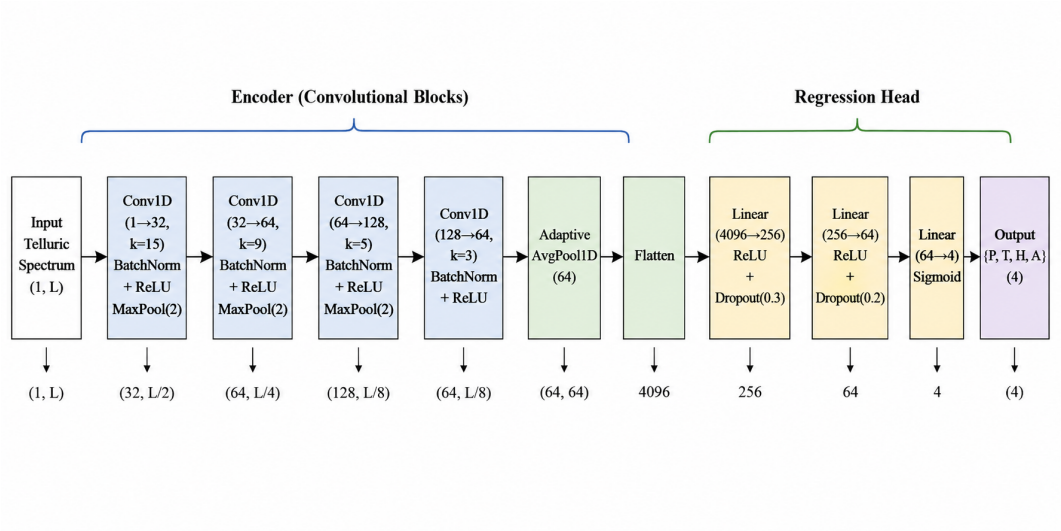

Figure 4.5: 1D-CNN. The encoder extracts hierarchical spectral features through four convolutional blocks. The regression head maps the pooled representation to $\{P, T, H, A\}$

.

4.4.3 Baseline comparisons

Before finalizing the 1D-CNN model, two simpler models were evaluated for this same task. These were a fully connected MLP (three hidden layers of $512 \rightarrow 256 \rightarrow 128$ units with batch normalization and dropout) and a Random Forest regressor (300 trees, $\sqrt{d}$ features per split). Both of these models were trained separately on Chunk 1 and Chunk 2. Their performance was significantly poorer compared to the adopted 1D-CNN on both chunks. Therefore they were discontinued for the combined Chunk 1 and Chunk 2 dataset and the subsequent 20,000 training run. Results are reported in Section 5.2.

4.4.4 Training

The CNN was trained on a synthetic dataset of Telfit generated telluric spectra with known (P, T, H, A) labels. Two dataset sizes were used: $\sim$10,000 spectra and then a run on an expanded set of $\sim$20,000 spectra to assess the effect of training data size and expanded parameter space on the output.

Each dataset was split 80/20 into training and test sets. Separate models were trained on Chunk 1, Chunk 2 and the combined Chunk 1 and Chunk 2. Adam

optimizer was used with mean-squared-error loss, a learning rate of 10^{-3} with a batch size of 128. Learning rate scheduler was used that halved the learning rate after every 30 epochs for total 100 epochs. The model with the lowest validation loss (best model) was retained.

4.4.5 Generating estimated telluric spectrum

The predicted parameter vector $\hat{\theta} = (\hat{P}, \hat{T}, \hat{H}, \hat{A})$ is inverse scaled back to physical units to match the original parameter range and then it is passed to Telfit synthetic telluric spectrum generator that runs a full line-by-line radiative transfer calculation at that atmospheric state and returns a clean synthetic telluric transmittance spectrum $\hat{T}(\lambda)$. This is the estimated Step 3 output, which is used in Step 4 to correct the observed spectrum. Figure 5.1 compares $\hat{T}(\lambda)$ with the observed Step 2 telluric template.

4.5 Step 4 - Obtaining telluric corrected spectrum

Observed spectrum is elementwise divided by the estimated telluric spectrum to finally obtain the telluric corrected spectrum.

$$\hat{F}_{\star}(\lambda) = \frac{F_{\mathrm{obs}}(\lambda)}{\hat{T}(\lambda)}. \tag{4.6}$$

where $F_{\mathrm{obs}}(\lambda)$ is the stitched normalized observed flux at wavelength λ, $\hat{T}(\lambda)$ is the estimated telluric transmittance and $\hat{F}_{\star}(\lambda)$ is the resulting telluric corrected flux at λ. The telluric corrected spectrum is then compared against another telluric corrected spectrum (Reference) of the same target star. Since the reference corrected spectrum was already BERV corrected, a reverse BERV correction was performed on it before comparing it with our obtained telluric corrected spectrum.

4.6 Bridging the real to synthetic gap

The 1-D CNN in Step 3 has only clean Telfit generated spectra (synthetic). But during the inference it is fed observed Step 2 telluric spectrum (real). The observed telluric spectrum carries masked residuals, continuum imperfections, and instrumental noise

which the synthetic spectrum doesn't have. This mismatch creates a problem of the pipeline and I tried two unsupervised domain-adaptation strategies to solve it.

4.6.1 CycleGAN translation (Cycle-StarNet style)

Following the idea behind Cycle-StarNet [10], a 1D CycleGAN [9] was trained with two generator-discriminator pairs. Generator $G_{s\to r}$ maps clean synthetic Telfit spectra into the observed Step 2 real domain telluric spectra while $G_{r\to s}$ does the reverse.

Both generators use the same architecture. The initial convolutional layer extracts low-level spectral features, followed by two downsampling layers that compress the spectral representation. The compressed features are then processed by six residual blocks, which helps it to learn the transformations to move between the synthetic and real domains. Two upsampling layers reconstruct the spectrum and a final convolutional layer with tanh activation function produces the output spectrum. Instance normalization is applied after each convolutional layer. Further reflection padding is used to minimize the edge artifacts near the boundaries of the spectrum.

The two discriminators follow the PatchGAN design. Instead of judging the entire spectrum at once, each discriminator evaluates a series of local spectral segments to determine whether it is real or generated. This encourages the generators to reproduce realistic telluric absorption features at local scale while preserving the overall spectral structure.

Both the generator and discriminator pairs were trained using the LSGAN objective [12] with separate Adam optimizers ($\eta_G = 2 \times 10^{-4}$, $\eta_D = 1 \times 10^{-4}$). The synthetic domain is represented by the 20,000 Telfit generated synthetic spectra. The real domain is represented by 41 observed telluric spectra. Cycle consistency weight was $\lambda_{\text{cyc}} = 10$ and identity weight was $\lambda_{\text{idt}} = 5$. A replay buffer was employed to stabilize adversarial updates. Training was performed for 100 epochs with a batch size of 32.

At inference, the observed Step 2 telluric spectrum is fed into the $G_{r\to s}$. It produces a cleaned synthetic spectrum. This translated cleaned synthetic spectrum can then be passed directly to the 1D-CNN described in Section 4.4.2 which then predicts four atmospheric parameters (P, T, H, A). It later feeds it into Telfit synthetic spectra generator to generate an estimated telluric spectrum for Step 4. Otherwise this translated cleaned synthetic spectrum itself can be used directly as the estimated telluric spectrum in Step 4, bypassing parameter retrieval from Telfit (eliminating the

slower process).

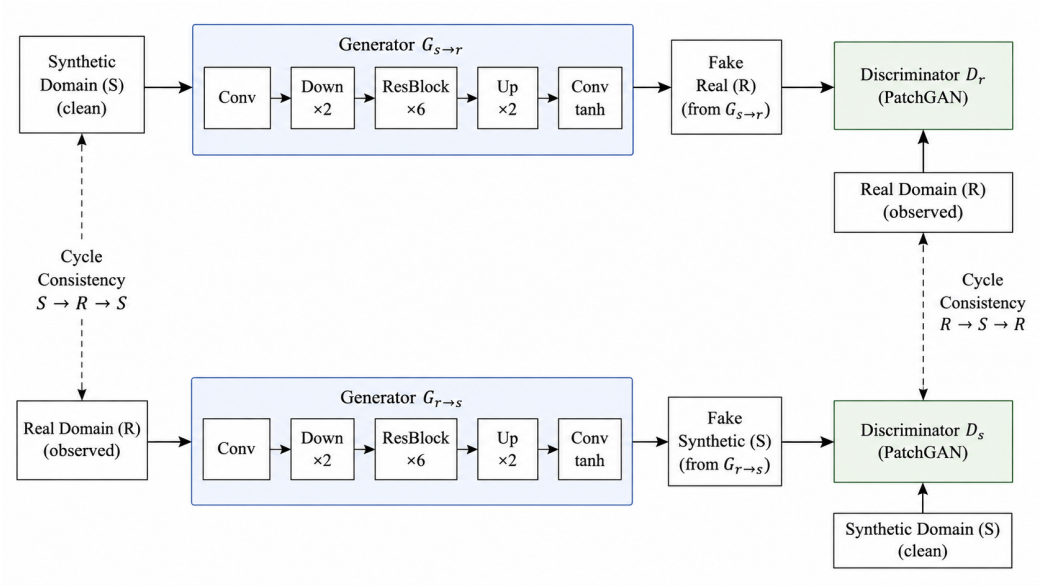

Figure 4.6: 1D CycleGAN. Two ResNet generators translate between the synthetic and real telluric domains. Two PatchGAN discriminators critique local patches. Cycle-consistency ties them.

4.6.2 Domain-adversarial auto-encoder (DAAE)

Using [11] [13], a domain-adversarial auto-encoder was trained to bridge the real-to-synthetic domain gap.

The model has a shared 1D-CNN encoder G_f connected to three heads: a decoder G_{dec} that reconstructs the input spectrum, a regressor G_y that predicts the atmospheric parameters $\{P, T, H, A\}$ and a domain classifier G_d that distinguishes synthetic from real spectra. G_d is connected through a gradient reversal layer (GRL). The GRL passes the signal unchanged in the forward direction but flips the sign of the gradient during backpropagation. In this manner the encoder gets trained to produce features that are useful for G_y yet carry no information that G_d can exploit to distinguish the two domains.

The encoder consists of four convolutional blocks followed by InstanceNorm and ReLU. The kernel size changes from 15, 9, 5 to 3. The number of channels grows from 32 to 128. Three max-pool layers reduce the spectral length by a factor of eight.

The decoder reverses this with three transpose convolution blocks, followed by a

convolution and a sigmoid output layer. G_d is kept intentionally small. It consists of only one hidden layer of 32 neurons with Dropout(0.5). This is because with only 34 observed telluric spectra (real) a larger classifier will easily memorize individual observations within a few epochs. This would lead to collapse of the adversarial gradient which can further halt the domain adaptation.

The model is trained on two domains simultaneously. The *source domain* consists of the same ∼20,000 synthetic Telfit transmittance spectra that was used to train the 1-D CNN with known labels (P, T, H, A). The synthetic set was split 90/10 into 18,000 training and 2,000 validation spectra. The *target domain* consists of the 34 observed telluric spectra that carry no labels. Now as the two domains are severely imbalanced, real spectra are cycled continuously using an infinite sampler so that every training batch contains examples from both domains. The model is trained by minimizing the combined loss:

$$\mathcal{L} = \lambda_{\text{task}}\,\mathcal{L}_{\text{task}} + \lambda_{\text{recon}}\,\mathcal{L}_{\text{recon}} + \lambda_{\text{floor}}\,\mathcal{L}_{\text{floor}} + \lambda\,\mathcal{L}_{\text{dom}}. \tag{4.7}$$

Task loss: $\mathcal{L}_{\text{task}}$ is the MSE loss between the predicted parameters $(\hat{P}, \hat{T}, \hat{H}, \hat{A})$ and the known labels on synthetic spectra.

Reconstruction loss: $\mathcal{L}_{\text{recon}}$ penalises errors between the decoder output $\hat{T}_i$ and the input spectrum T_i at each wavelength pixel i:

$$\mathcal{L}_{\text{recon}} = \frac{1}{L}\sum_{i=1}^{L} \underbrace{\left[1 + \beta\,|T_i - 1|\right]}_{\text{depth weight}} |\hat{T}_i - T_i|, \qquad \beta = 4. \tag{4.8}$$

The depth weight $1 + \beta|T_i - 1|$ is large at absorption lines (where $T_i \ll 1$) and close to 1 at flat regions (where $T_i \approx 1$). This ensures matching of deep telluric features while continuum noise gets smoothened.

Floor loss: $\mathcal{L}_{\text{floor}}$ enforces the physical constraint that the reconstructed telluric transmittance should not fall below the observed flux at deep absorption regions:

$$\mathcal{L}_{\text{floor}} = \frac{\sum_i \text{ReLU}(x_i - \hat{T}_i) \cdot \mathbf{1}[x_i < 0.9]}{\sum_i \mathbf{1}[x_i < 0.9] + 1}, \tag{4.9}$$

where x_i is the observed flux at wavelength point i. The indicator $\mathbf{1}[x_i < 0.9]$ restricts this penalty to deep telluric dip regions only, leaving the continuum unconstrained so

the model does not collapse to simply copying the input.

Domain loss: $\mathcal{L}_{\text{dom}}$ is the binary cross entropy loss of G_d classifying each spectrum as synthetic (0) or real (1).

The loss weights are $\lambda_{\text{task}} = 1$, $\lambda_{\text{recon}} = 5$, $\lambda_{\text{floor}} = 1$. High reconstruction weight of 5 keeps the decoder grounded throughout training. The adversarial strength λ is not fixed. It begins at zero and increases gradually as training progresses. It uses :

$$\lambda_p = \alpha\left(\frac{2}{1+e^{-\gamma p}} - 1\right), \qquad \alpha = 0.5,\ \gamma = 10, \tag{4.10}$$

where $p \in [0, 1]$ is the fraction of training completed. This allows the model to first learn accurate spectrum reconstruction and atmospheric-parameter prediction before proceeding to domain-invariant feature learning. All components were trained jointly using Adam optimizer with a learning rate of 10^{-3} and a weight decay of 10^{-5}. Cosine annealing schedule was used and the model was trained for 120 epochs.

At inference, the observed Step 2 telluric spectrum is passed through the shared encoder G_f, which strips domain-specific artifacts (noise, continuum ripple) via the trained GRL to produce domain-invariant features. The decoder G_{dec} then reconstructs a clean synthetic-style telluric spectrum $\hat{x}_s$ from these features. Because the decoder was trained on synthetic spectra, it naturally outputs a flat continuum at 1.0 between absorption lines. At the same time, the regressor head forces the latent representation to retain the atmospheric physics, so dip depths from the real observation are preserved. This translated spectrum $\hat{x}_s$ can then be passed to the 1D-CNN (Section 4.4.2), which predicts (P, T, H, A) and then the Telfit synthetic spectra generator generates a refined telluric model for Step 4. Otherwise $\hat{x}_s$ can directly be used in Step 4.

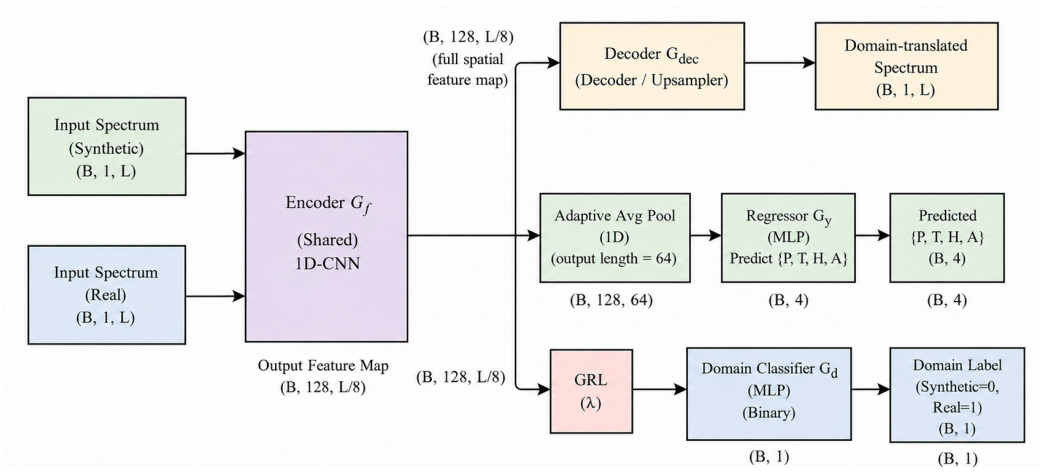

Figure 4.7: Domain-adversarial auto-encoder. A shared encoder feeds a reconstruction decoder, a $\{P, T, H, A\}$ regressor, and a domain classifier behind a gradient-reversal layer, producing domain-invariant features that close the synthetic-to-real gap.

Chapter 5

Experiments and Results

5.1 Payne emulator

The Payne emulator was trained twice. First it was trained on the 171 original BT-Settl spectra. Later it was trained on the expanded set of 14,266 training spectra that was produced by grid interpolation. In both cases the test set consists of the same 43 spectra that were previously separated out prior to interpolation. Table 5.1 reports the test-set metrics for all four configurations.

Table 5.1: *Payne* emulator performance on the held-out test set (43 BT-Settl spectra) for both spectral chunks. R^2_{global} is computed over all pixels jointly; R^2_{star} is the mean per-spectrum R^2.

Training Set	Chunk	R^2_{global}	R^2_{star}	MAE
171	755–775 nm	0.897	0.871	0.0346
171	830–850 nm	0.859	0.839	0.0276
14,266	**755–775 nm**	**0.965**	**0.962**	**0.0201**
14,266	**830–850 nm**	**0.984**	**0.974**	**0.0113**

Grid expansion improved results across both chunks. For Chunk 1 (755–775 nm) the R^2_{global} improved from 0.897 to 0.965 and MAE dropped by around 42%. While for Chunk 2 (830–850 nm) the R^2_{global} improved from 0.859 to 0.984 and MAE dropped by 59%. The train test R^2 gap remains below 0.035 in all cases that indicates no significant overfitting. Thus the expanded Payne emulator reconstructs the stellar continuum shape and major absorption line positions.

5.2 Performance on Atmospheric Parameter Estimation

5.2.1 Chunkwise Comparison

Table 5.2 reports the test set R^2 for all three models. These models were trained separately on Chunk 1 and Chunk 2 using the Telfit Synthetic Spectra Generator generated $\sim$10,000 synthetic dataset. The MLP fails to predict entirely on humidity in Chunk 1 ($R^2 = 0.00$) and gives negative R^2 (indicating worse prediction than mean) for pressure and airmass in Chunk 2. The Random Forest model severely overfits. It gives train $R^2 \approx 1$ for every atmospheric parameter while test scores for pressure (-2.20, $+0.52$) and airmass ($+0.89$, -0.92) are also not consistent across both chunks. The 1D-CNN produces stable train test gaps (≤ 0.004 for all parameters, Table 5.3) and the highest test R^2 among all the parameters. However it fails to predict the correct humidity on Chunk 1. Although all three models fail to do that , as the variation in humidity parameter is negligible across Chunk 1.

Table 5.2: Test-set R^2 for MLP, RF, and 1D-CNN trained on the $\sim$10,000 dataset and evaluated separately on Chunk 1 (755–775 nm) and Chunk 2 (830–850 nm). Negative values indicate predictions worse than the mean value.

Model	**Chunk**	**Pressure**	**Temperature**	**Humidity**	**Airmass**
MLP	1	0.426	0.941	0.000	0.813
MLP	2	-3.279	0.865	0.708	-1.169
RF	1	-2.200	0.862	0.959	0.886
RF	2	0.520	0.978	0.654	-0.922
1D-CNN	1	**0.883**	**0.943**	-0.005	**0.962**
1D-CNN	2	**0.958**	**0.989**	**0.960**	**0.791**

Chunk 1 actually covers the O_2 A-band (755–775 nm) with negligible water vapor absorption. Thus the flux varies very little although the humidity is varied. Chunk 2 (830–850 nm) contains significant water vapor absorption lines and the 1D-CNN successfully predicts the humidity at $R^2 = 0.960$.

Table 5.3: Train–test R^2 gap for the 1D-CNN on individual chunks (10,000 dataset), confirming absence of overfitting.

Chunk	Split	Pressure	Temperature	Humidity	Airmass
1	Train	0.879	0.943	0.004	0.962
1	Test	0.883	0.943	−0.005	0.962
2	Train	0.958	0.989	0.962	0.819
2	Test	0.958	0.989	0.960	0.791

5.2.2 Combined-Chunk 1D-CNN: effect of training set size

As the 1D-CNN outperformed the MLP and RF, they were discontinued from further experimentation. Table 5.4 compares the 1D-CNN trained on the combined Chunk 1 and Chunk 2 dataset at ∼10,000 and ∼20,000 spectra. The 20,000 dataset uses an expanded parameter grid (Table 2.2).

Table 5.4: Test-set metrics for the 1D-CNN on the combined Chunk 1 and Chunk 2 dataset at two different training sizes. The 20,000 run uses an expanded parameter grid.

Metric	Train set	Pressure (hPa)	Temp. (K)	Humidity (%)	Airmass
R^2	10,000	0.607	0.985	0.979	0.984
R^2	20,000	**0.976**	**0.997**	**0.992**	**0.996**
MSE	20,000	4.417	0.269	5.928	0.00105
RMSE	20,000	2.102	0.518	2.435	0.0324

Doubling the training set and expanding the parameter grid raised the pressure R^2 from 0.607 to 0.976. This closed the gap that persisted on the single chunk models. Temperature, humidity and airmass already reached $R^2 = 0.97$ at 10,000. It improved further to ∼0.99 for 20,000 dataset. This improvement is also reflected in the low MSE and RMSE values of the 20,000 model, with the RMSE for pressure, temperature, humidity and airmass at 2.102 hPa, 0.518 K, 2.435 % and 0.0324 respectively. Furthermore, the train-test gaps also remain ≤0.003 (Table 5.5) which further indicates the model's generalization ability and ability to overcome overfitting.

Table 5.5: Train test R^2 gap for the combined 20,000 1D-CNN model.

Split	**Pressure**	**Temperature**	**Humidity**	**Airmass**
Train	0.978	0.997	0.992	0.996
Val	0.977	0.997	0.992	0.996
Test	0.976	0.997	0.992	0.996
Gap (Train Test)	0.002	0.000	0.000	0.000

The combined Chunk 1 and Chunk 2 model trained on 20,000 spectra is therefore adopted for all subsequent evaluations.

5.3 CycleGAN Performance

The CycleGAN was trained for 100 epochs. During training the real domain discriminator (D_r) collapsed rapidly at an alarming rate. The D_r loss dropped from 0.2428 to 0.0080 within only 8 epochs and after that always remained close to zero. In contrast, the synthetic-domain discriminator (D_s) remained stable at $\sim$0.20 throughout the whole training phase.

The adversarial loss (G_{adv}) steadily increased from 0.8500 to 1.3310 at the end of all epochs. However the cycle consistency loss (G_{cyc}) decreased from 1.5732 to 0.3716.

But as the D_r failed early in training the real to synthetic generator (G_{r2s}) no longer received meaningful adversarial feedback. This led to poor synthetic reconstruction of real spectra although the cycle consistency loss reduced throughout the whole training phase.

5.4 DAAE Performance

The DAAE was trained for 120 epochs with the adversarial weight λ ramping from 0.02 to 0.50 over the course of training. The task loss decreased steadily from 0.049 at epoch 1 to 0.017 at epoch 120. The reconstruction loss too fell from 0.056 to 0.010 at the end of 120 epochs. The domain loss dropped from 0.027 to 0.001, indicating that the encoder won the adversarial game: the shared encoder G_f learned to produce features from which the domain classifier could no longer distinguish synthetic from real spectra.

All three losses improved simultaneously, confirming that domain-invariant feature learning, atmospheric parameter prediction and spectrum reconstruction proceeded without conflicting with each other. As a result, the decoder produces clean synthetic-style output from real observations. Validation task loss stabilized near 0.0226 and showed no sign of overfitting. Further detailed evaluation against the reference telluric corrected spectrum is available in Section 5.6.

5.5 Comparison of Step 2 output with all estimated telluric spectra

Figure 5.1 overlays estimated telluric spectrum produced by all model variations with the observed Step 2 telluric spectrum. A small region (7665–7675 Å) is taken.

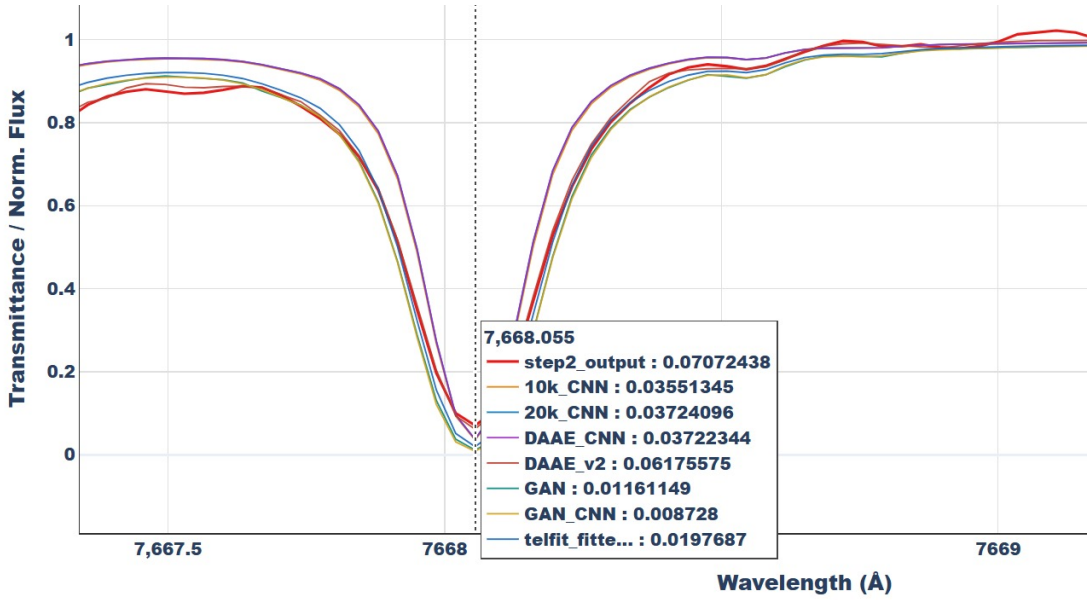

Figure 5.1: Overlay plot of Step 2 output against all variations. DAAE (deep orange) almost replicates the Step 2 telluric dips while maintaining smoothness in stellar regions. CNN based methods and Telfit fitted predict shallower cores. And the CycleGAN outputs (GAN, GAN_CNN) are severely shallow which further confirms discriminator collapse problem.

DAAE is able to perfectly replicate the Step 2 telluric depths. At 7668 Å the dip depth is 0.071 for Step 2 output while it is 0.062 for DAAE, thus they are very

close. Furthermore, at other non telluric regions it preserves the smoothness of the continuum that is usually observed in synthetic telluric spectrum. Both the CNN based methods (10,000, 20,000) and the DAAE output fed to CNN predict slightly shallower cores in the range 0.035 to 0.037. The Telfit fitted spectrum sits at 0.020, shallower than the CNN methods but deeper than the CycleGAN outputs. Both the CycleGAN variants (GAN, GAN_CNN that is CycleGAN output fed to CNN) predict severely shallow depths from 0.009 to 0.012. This further confirms discriminator collapse.

5.6 Step 4: Results

5.6.1 Data splits

To avoid confusion, the data splits used across the pipeline are summarized here. The synthetic dataset consists of ∼20,000 Telfit-generated transmittance spectra. A 90/10 split produces 18,000 training and 2,000 validation spectra. This split is shared by both the 1D-CNN (which uses synthetic spectra exclusively) and the DAAE source domain. The real dataset consists of 41 observed telluric spectra extracted from CARMENES 2018 observations. For the DAAE, 34 of these serve as the unlabelled target domain during training and the remaining 7 are held out as unseen test stars for generalization evaluation.

5.6.2 Evaluation against telluric corrected reference

Seven model variations were evaluated against the already available reference spectrum on the basis of two losses. They are : the mean squared error (MSE) and the mean absolute error (MAE). Loss is thus computed between Step 4 telluric corrected flux $\hat{F}_\star(\lambda)$ and the reference $F_{\mathrm{ref}}(\lambda)$ over all 11,432 wavelength points (that is for the merged chunks).

Single representative star. Table 5.6 reports the results on a single representative observation (J19346+045).

Table 5.6: Evaluation of seven variations against reference spectrum on a single representative observation (J19346+045). MSE and MAE are computed over the merged Chunk 1 and Chunk 2 and the best result is highlighted.

Method	**MSE**	**MAE**
DAAE (direct)	$\mathbf{1.19 \times 10^{-2}}$	**0.0381**
20,000 CNN	1.05×10^{1}	0.2450
DAAE + CNN	1.05×10^{1}	0.2452
10,000 CNN	4.48×10^{1}	0.4246
CycleGAN + CNN	1.18×10^{7}	114.01
CycleGAN (direct)	1.32×10^{7}	49.03

DAAE generalization across 7 unseen test stars. To evaluate generalization, the DAAE trained on 34 observed spectra was tested on 7 held-out stars that were never seen during training. Table 5.7 reports per-star results. The overall metrics are computed across all 80,024 wavelength points (11,432 per star $\times$ 7 stars).

Table 5.7: Per-star DAAE (direct) evaluation against 2023 reference on 7 unseen test stars.

Observation	**MSE**	**MAE**
J19346+045	0.0119	0.0381
J18580+059	0.0092	0.0424
J11026+219	0.0100	0.0432
J23492+024	0.0150	0.0589
J08161+013	0.0170	0.0657
J17355+616	0.0348	0.0721
J09428+700	0.0310	0.0913
Overall	**0.0184**	**0.0588**

Four out of seven stars achieve MAE < 0.06, with the best MAE of 0.038 on J19346+045. The consistent performance across different stars, observation dates and atmospheric conditions confirms that the DAAE has learned a general domain translation rather than memorizing the 34 training observations.

5.6.3 Corrected spectra

Figure 5.2 overlays the Step 4 telluric corrected spectra of all seven model variations with the reference spectrum. It also includes the telluric contaminated observed spectra. A small region is taken for this purpose.

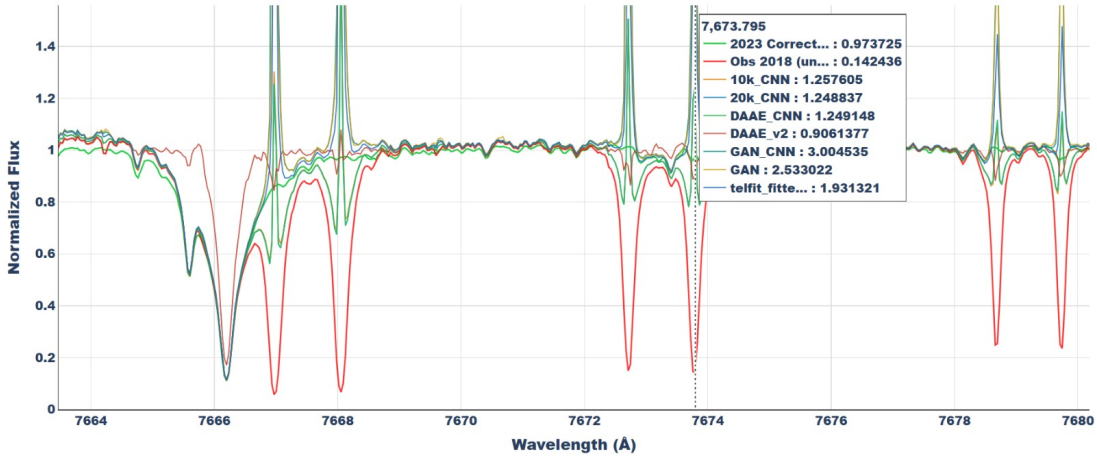

Figure 5.2: Step 4 telluric corrected spectra from all seven model variations are overlaid on the reference spectrum (green) and the telluric contaminated observed spectrum (red). DAAE (deep orange) replicates the reference most closely at deep telluric lines. The Telfit fitted spectrum over-corrects with flux reaching 1.93. The CycleGAN variants (GAN, GAN_CNN) produce spikes with flux exceeding 2.5.

5.6.4 Discussion

DAAE direct: The DAAE output used directly for Step 4 achieves the lowest MAE error (MAE 0.038 on the representative star, overall MAE 0.059 across 7 unseen test stars). At the telluric line (7673.8 Å) it gives the telluric corrected flux as 0.906 against the reference value of 0.974 which is actually the best approximation against all the other model variations. This indicates that the DAAE preserves telluric line depths of Step 2 output. Furthermore, it also preserves the noise of the telluric contaminated observed spectrum which is necessary for stellar parameter estimation.

CNN: The 20,000 CNN model with (MAE 0.245) outperforms the 10,000 CNN with (MAE 0.425) by a factor of 1.7. This confirms that the larger synthetic training set with wider parameter range is able to perform better on real observed telluric

spectrum. Although both of the variants overcorrect the telluric line depths as a result direct pointwise division generates spikes (flux ≈ 1.25 at 7673.8 Å).

DAAE + CNN. DAAE+CNN (MAE 0.245) performs nearly identically to the 20,000 CNN. The DAAE reduces the domain gap but the CNN was trained purely on clean synthetic Telfit spectra. So it has never seen real observed telluric spectrum. Therefore the translated spectrum still appears to be out of distribution to the CNN model and thus there remains a significant error.

Telfit fitted. The Telfit fitted spectrum over-corrects at deep telluric lines, producing flux values of 1.93 at 7673.8 Å against the reference value of 0.974. This is because the Telfit forward model, while physically grounded, predicts slightly shallower telluric depths (0.020 at 7668 Å) compared to the observed Step 2 telluric spectrum (0.071). Division by this shallower estimate amplifies the corrected flux.

CycleGAN. Both CycleGAN variants produce large errors (MAE 49 and 114, MSE $\sim 10^7$). Severe underestimation of telluric depths amplifies the flux values at the telluric line depths. Thus it produces huge emission spikes of 2.53 and 3.00 at 7673.8 Å against the reference value of 0.974. These results are a direct consequence of discriminator collapse where only 34 real spectra are available against 20,000 synthetic. The real domain discriminator (D_r) memorized the real domain within eight epochs. This further led to mode collapse in the generator.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This work presents a telluric correction pipeline for correcting CARMENES M-dwarf telluric contaminated observed spectra. This proposed pipeline requires no standard star or ground truth corrected spectrum. Seven different model variants were evaluated against a reference spectrum.

The DAAE used directly as the estimated telluric spectrum for Step 4 achieved the best performance (MAE 0.038 on a representative star, overall MAE 0.059 across 7 unseen test stars). The consistent performance across different stars, observation dates and atmospheric conditions confirms the generalization ability of the DAAE. The 20,000 CNN outperforms the 10,000 CNN by a factor of 1.7 which confirms that wider parameter space coverage reduces error. The DAAE+CNN result reveals that domain adaptation and retrieval cannot be treated as two independent stages. The CNN trained purely on synthetic spectra thus cannot cope with the translated domain produced by DAAE. The Telfit fitted spectrum over-corrects at deep telluric lines due to shallower predicted telluric depths. The severe class imbalance between 41 real and 20,000 synthetic spectra causes discriminator collapse. This indeed makes the CycleGAN variants the worst performing among all other variants.

These results demonstrate that the proposed pipeline successfully replicates the correction quality of the telluric corrected reference spectrum. Unlike traditional tools such as Telfit and Molecfit which require a large amount of time for fitting this pipeline significantly reduces the time. As a result, the pipeline provides a significantly faster solution for telluric correction of M-dwarf spectra. However generalization to the full CARMENES VIS region remains an open question and the subject of future work.

6.2 Challenges

Stellar masking: Identifying stellar absorption regions required manually consulting both BT-Settl synthetic stellar spectra and Telfit transmittance spectra. This process was time consuming and error prone. Heuristic and automated processes although produced results but they were not accurate enough and masking accuracy directly determines the quality of everything downstream. Further scaling this proposed pipeline to the entire VIS or NIR regions would require huge amount of manual effort if accuracy needs to be preserved.

Synthetic-to-real domain gap. The retrieval CNN was trained purely on clean synthetic spectra but at inference it received observed telluric spectra. Bridging this distribution shift was the main challenge of this work. The DAAE partially solves this issue and produces good results but there is still a gap in certain regions.

6.3 Future Work

Automated stellar masking: The current masking step is largely manual and is therefore time consuming. In future a classifier can be trained to predict each region and then the stellar features can be masked out automatically. This will significantly reduce human effort and make this proposed pipeline scalable to larger datasets and broader spectral regions.

Expand the observed dataset: CycleGAN failure was directly caused by the severe imbalance between 20,000 synthetic and 41 real spectra. A larger real dataset can reduce this class imbalance. This will allow the CycleGAN to train properly and avoid discriminator collapse.

Widen the synthetic parameter grid: The jump in CNN performance on 10,000 versus the 20,000 training set confirms that broader parameter space coverage directly improves atmospheric parameter retrieval accuracy. The current Telfit grid still covers a limited part of the atmospheric parameter space. So further expanding the parameter range can further improve CNN performance and its overall generalization ability.

Jointly train DAAE and CNN: The DAAE+CNN result shows poor performance. In future joint training of the DAAE and CNN can be applied. This will enable the CNN to adapt to the translated distribution.

Expansion across full VIS and NIR regions: The pipeline currently operates on two spectral chunks only. In future this proposed pipeline can be extended to the full CARMENES VIS region and subsequently to the NIR region (960–1710 nm), where telluric contamination is greater.

Replace Telfit synthetic spectra generator with a deep learning forward model: At present each call to the Telfit synthetic spectra generator takes approximately 15 minutes. In future this can be directly replaced with a neural network where on feeding the atmospheric parameters it gives a synthetic telluric spectrum. This will significantly reduce the per observation correction time and make this scalable.

References

[1] A. Quirrenbach et al., "CARMENES: An Overview Six Months after First Light," *Proc. SPIE*, vol. 9908, p. 990812, 2016.

[2] W. D. Vacca, M. C. Cushing, and J. T. Rayner, "A Method of Correcting Near-Infrared Spectra for Telluric Absorption," *PASP*, vol. 115, no. 805, p. 389, 2003.

[3] A. Smette et al., "Molecfit: A general tool for telluric absorption correction," *Astronomy & Astrophysics*, vol. 576, A77, 2015.

[4] K. Gullikson, S. Dodson-Robinson, and A. Kraus, "Correcting for Telluric Absorption: Methods, Case Studies, and the `Telfit` Code," *The Astronomical Journal*, vol. 148, no. 3, p. 53, 2014.

[5] F. Allard et al., "Models of very-low-mass stars, brown dwarfs and exoplanets," *Phil. Trans. R. Soc. A*, vol. 370, no. 1968, p. 2765, 2012.

[6] Y.-S. Ting et al., "The Payne: Self-consistent Machine Learning of Stellar Parameters and Spectra," *The Astrophysical Journal*, vol. 879, no. 2, p. 69, 2019.

[7] É. Artigau et al., "Telluric-line subtraction in high-accuracy velocimetry: A PCA-based approach," *Proc. SPIE*, vol. 9149, 914905, 2014.

[8] S. Fabbro et al., "An Application of Deep Learning in the Analysis of Stellar Spectra," *MNRAS*, vol. 475, no. 3, p. 2978, 2018.

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," in *Proc. ICCV*, 2017, pp. 2223–2232.

[10] T. O'Briain et al., "Cycle-StarNet: Bridging the Supervised and Unsupervised Domain Gap in Stellar Spectra," *The Astrophysical Journal*, vol. 906, no. 2, p. 130, 2021.

[11] Y. Ganin et al., "Domain-Adversarial Training of Neural Networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[12] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least Squares Generative Adversarial Networks," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2794–2802.

[13] H. S. Mavikumbure, V. Cobilean, C. S. Wickramasinghe, B. J. Varghese, T. Pennington, and M. Manic, "DAdAE: Domain Adversarial Autoencoder Based In-Vehicle CAN Anomaly Detection," in *Proc. IECON 2023 – 49th Annual Conference of the IEEE Industrial Electronics Society*, Singapore, 2023, pp. 1–7.