Galaxy Spectra neural Network (GaSNet). II. Using Deep Learning for Spectral Classification and Redshift Predictions

Fucheng Zhong, Nicola R. Napolitano, Caroline Heneka, Rui Li, Franz Erik Bauer, Johan Comparat, Young-Lo Kim, Jens-Kristian Krogager, Marcella Longhetti, Jonathan Loveday, et al.

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **532,** 643–665 (2024)
Advance Access publication 2024 June 27

https://doi.org/10.1093/mnras/stae1461

Galaxy Spectra neural Network (GaSNet). II. Using deep learning for spectral classification and redshift predictions

Fucheng Zhong,[1,2] Nicola R. Napolitano ,[1,2]★ Caroline Heneka ,[3] Rui Li ,[4] Franz Erik Bauer,[5,6,7] Nicolas Bouche ,[8] Johan Comparat,[9] Young-Lo Kim ,[10] Jens-Kristian Krogager,[11] Marcella Longhetti,[12] Jonathan Loveday ,[13] Boudewijn F. Roukema ,[14,15] Benedict L. Rouse,[5] Mara Salvato ,[16] Crescenzo Tortora ,[17] Roberto J. Assef,[18] Letizia P. Cassarà,[19] Luca Costantin,[20] Scott M. Croom ,[21] Luke J M Davies,[22] Alexander Fritz,[23] Guillaume Guiglion,[24,25,26] Andrew Humphrey,[27,28] Emanuela Pompei,[29] Claudio Ricci,[18,30] Cristóbal Sifón,[31] Elmo Tempel[32,33] and Tayyaba Zafar[34]

Affiliations are listed at the end of the paper

Accepted 2024 May 31. Received 2024 April 14; in original form 2023 November 7

ABSTRACT

The size and complexity reached by the large sky spectroscopic surveys require efficient, accurate, and flexible automated tools for data analysis and science exploitation. We present the Galaxy Spectra Network/GaSNet-II, a supervised multinetwork deep learning tool for spectra classification and redshift prediction. GaSNet-II can be trained to identify a customized number of classes and optimize the redshift predictions. Redshift errors are determined via an ensemble/pseudo-Monte Carlo test obtained by randomizing the weights of the network-of-networks structure. As a demonstration of the capability of GaSNet-II, we use 260k Sloan Digital Sky Survey spectra from Data Release 16, separated into 13 classes including 140k galactic, and 120k extragalactic objects. GaSNet-II achieves 92.4 per cent average classification accuracy over the 13 classes and mean redshift errors of approximately 0.23 per cent for galaxies and 2.1 per cent for quasars. We further train/test the pipeline on a sample of 200k 4MOST (4-metre Multi-Object Spectroscopic Telescope) mock spectra and 21k publicly released DESI (Dark Energy Spectroscopic Instrument) spectra. On 4MOST mock data, we reach 93.4 per cent accuracy in 10-class classification and mean redshift error of 0.55 per cent for galaxies and 0.3 per cent for active galactic nuclei. On DESI data, we reach 96 per cent accuracy in (star/galaxy/quasar only) classification and mean redshift error of 2.8 per cent for galaxies and 4.8 per cent for quasars, despite the small sample size available. GaSNet-II can process ∼40k spectra in less than one minute, on a normal Desktop GPU. This makes the pipeline particularly suitable for real-time analyses and feedback loops for optimization of Stage-IV survey observations.

Key words: methods: data analysis – techniques: spectroscopic – surveys – software: development – galaxies: distances and redshifts.

1 INTRODUCTION

With the upcoming all-sky spectroscopic survey infrastructures, including the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration 2022), 4-metre Multi-Object Spectroscopic Telescope (4MOST; de Jong et al. 2019), Multi-Object Optical and Near-infrared Spectrograph (MOONS; Cirasuolo et al. 2020), and considering also the slitless spectroscopic capabilities of the space-based missions like *Chinese Space Station Telescope* (*CSST*; Zhan 2011) and *Euclid* (Laureijs et al. 2011), hundreds of millions of spectra will be acquired in the next half-decade. The first samples from DESI are already publicly available (DESI Collaboration 2023). To optimize the scientific outcome of these huge data sets, strategies to perform fast, efficient, and, most of all, accurate automated analyses have become mandatory. Machine learning (ML) provides a large variety of efficient solutions to achieve this goal. We have already demonstrated that convolutional neural network (CNN) models can be very effective in classifying spectra for specific tasks like the search for strong galaxy–galaxy lenses (GaSNet; Zhong, Li & Napolitano 2022), showing superior efficiency and flexibility compared to traditional methods [e.g. principal component analysis (PCA) eigenspectra fitting; see Talbot et al. 2021].

Object classification and redshift prediction are the first steps to be performed by standard pipelines of spectroscopy observations. They provide basic information to be used for science applications. For instance, the separation of quiescent early-type galaxies, from the starburst emitting systems is fundamental for galaxy formation (Lehnert & Heckman 1996), while the classification of active galactic nuclei (AGN) is crucial to understanding the role of supermassive black holes (Fiore et al. 2017), and the identification of quasars (quasi-stellar objects, QSOs) is important for cosmological studies (Secrest et al. 2021). ML can be an efficient and practical alternative

★ E-mail: napolitano@mail.sysu.edu.cn

to traditionally automatic methods (Bolton et al. 2012; Hutchinson et al. 2016) to build entire ML-based parallel pipelines, similar to what is already done in astronomical imaging, where there have been enormous advances in recent years. Some examples of these latter applications are the galaxy morphology pipelines, like the one developed by Domínguez Sánchez et al. (2022) for SDSS-DR17 (Sloan Digital Sky Survey-Data Release 17), and the pipeline developed by Boucaud et al. (2020) for *Euclid*. ML can offer huge decreases in computational time and resources (Graff et al. 2014), while providing close to human-level classification results, for example, in the star/quasar separation (Busca & Balland 2018). This provides the chance to overcome the limits typically plaguing traditional classification methods in terms of computational resources, human intervention, limited real-time applications, scalability, etc. (Alzubaidi et al. 2021), thus giving us the opportunity to develop automatized ML-based tools (D'Isanto & Polsterer 2018; Parks et al. 2018; Makhija et al. 2019).

With respect to spectroscopy, a variety of automatic redshift prediction tools and pipelines have been developed using traditional methods, but relatively little has been done in terms of ML applications. Traditional codes, such as SPECTRO1D (SubbaRao et al. 2002) and REDMONSTER (Hutchinson et al. 2016), based on cross-correlation methods (Tonry & Davis 1979), or REDROCK (Lan et al. 2023), based on template fitting using a set of different PCA components (DESI Collaboration 2023), are some examples of such automated tools. They have been tested or successfully applied to larger scale spectroscopy surveys, generally requiring minimal human intervention. However, they are often time-consuming, for example, if the number of templates increases, or require an optimization of the first guess redshifts to maximize the accuracy. Furthermore, in low signal-to-noise ratio (SNR) situations, the performance of some of these tools can highly be degraded (e.g. because of an increasing failure rate, Bolton et al. 2012).

Deep learning (DL) based methods, instead, have the advantage of efficiency, scalability, and flexibility. Here, the applications to spectroscopy are yet at the pioneer level and limited to the search for strong gravitational lenses, Li et al. 2019), with only a DL tool previously tested to classify spectra and measure redshift (i.e. GaSNet, Zhong et al. 2022) yet with the specific goal of finding hidden strong lensing emissions in galaxy spectra. However, the first GaSNet is versatile enough to be adapted to answer most of the typical problems large sky surveys might need to face. In particular, it can easily perform tasks like real-time analysis for the detection of transients/peculiar objects, and still give a prediction of their redshift.

In this paper, we present a new DL tool that expands the capabilities of the former GaSNet to respond to the needs for upcoming spectroscopic surveys like 4MOST and DESI. DESI is expected to observe 30 million galaxies/AGN and 10 million stars. On the other hand, 4MOST will cover approximately 15 000 sq deg and observe more than 25 million targets. In particular, we design and test a full real-time pipeline based on DL that uses reduced 1D spectra as input to (1) classify spectra in a given number of subclasses; (2) predict the redshift; and (3) assign an error to the redshift. GaSNet-II is a DL-based tool for spectroscopy classification and redshift prediction which provides the probability of the type of spectrum and the object redshift with uncertainty. To train and test the pipeline we start from a catalogue from SDSS-DR16 (Jönsson et al. 2020) which provides a large number of classified spectra grouped into about 180 classes. This allows us to randomly select 13 subclass spectra from the SDSS-DR 16, each with more than 20 000 spectra. The 4MOST mock spectra (10 subclasses) and DESI early data release spectra (3 classes) are also randomly selected as additional data sets, to examine the flexibility and generality of the pipeline. In particular, the different properties of these three data sets will allow us to cover a large variety of classification situations from very specialized classifications for SDSS and 4MOST samples to a coarse-grained classification using DESI data.

The paper is organized as follows: in Section 2, SDSS data sets used for our analysis are introduced. In Section 3, we describe the ML models and our novel idea of building an ML pipeline. In Section 4, we present the training and testing results. In Section 5, we discuss the ML predicted results, including further improvements and perspectives for further ML pipelines. In the final Section 6, we draw some conclusions.

2 DATA

The main purpose of this paper is to find a DL-based method, to classify and predict the redshift of 1D spectra. As introduced above, we are interested in applying 'supervised' networks, based on labelled data. For the scope of this work, the main labels we need to start with are a 'class' and a 'redshift'. The generality of the tool depends on the number of classes we can separate from their spectral properties. While a basic separation can rely on a very coarse classification aiming to distinguish only stars, galaxies, and AGN/QSO (Pâris et al. 2017), for many science applications, one might be interested in a more detailed classification that distinguishes various star, galaxy, and AGN/QSO subclasses (Bundy et al. 2015; Yan et al. 2019). In this case, to best train any supervised tool we need data sets that can provide such kind of information. The ideal data set would be an observed sample of objects for which a qualitative/quantitative classification has been performed (Liu et al. 2019; Lyke et al. 2020). However, as an alternative, one can use mock data sets, where physically motivated templates of different galactic and extragalactic objects in different instrumental conditions (resolution, seeing, etc.) and covering a realistic range of intrinsic object properties (e.g. luminosity, colours, redshifts, kinematics, etc.), can mimic the data one is expected to collect for a given science program (e.g. via spectral synthesis; Cid Fernandes et al. 2005).

Below we describe the data we will use throughout the paper, covering the two typologies of training/test samples discussed above. In particular, as the observation-based data set, we use the SDSS-DR16 data set, which contains the most detailed classified subclass sample of sources available to date. As such, this will represent the reference data set around which we want to construct and benchmark our pipeline. Furthermore, to explore the possible application of GaSNet-II to upcoming stage-IV surveys, we use a customized mock catalogue, closely reproducing 4MOST observations (Driver et al. 2019; Helmi et al. 2019; Merloni et al. 2019; Swann et al. 2019; de Jong et al. 2019). Furthermore, we take advantage of the early data release of DESI (DESI Collaboration 2023), to perform a first test of the novel GaSNet-II version performances on a first Stage-IV survey data set. Notable for SDSS and DESI, the redshifts and classifications are not 100 per cent reliable (see e.g. Lyke et al. 2020; Alexander et al. 2023), which can potentially lead to deviations between the DL predictions and the pipeline results.

2.1 Reference data set: SDSS-DR16

SDSS-DR16 (Ahumada et al. 2020), contains around 0.44 million unique stars, 2.6 million galaxies, and 0.75 million quasars; all spectra are divided into three classes (star, galaxies, and QSOs), each one having a different number of subclasses for a total of 181 subclasses. Most of the subclasses comprise a number of spectra

Table 1. Some definitions and statistics of our reference data set from SDSS.

Column 1	2	3	4	5
class_subclass	Label	$\bar{z}$	$[z_{min}, z_{max}]$	$\overline{SNR}$
STAR_A0	0	–	–	26.2
STAR_F5	1	–	–	30.5
STAR_F9	2	–	–	34.9
STAR_G2	3	–	–	33.7
STAR_K1	4	–	–	32.8
STAR_K3	5	–	–	31.1
STAR_K5	6	–	–	31.0
GALAXY_nan	7	0.46	[0.00, 1.86]	5.82
GALAXY_AGN	8	0.21	[0.00, 0.57]	14.3
GALAXY_STARBURST	9	0.15	[0.00, 0.57]	9.78
GALAXY_STARFORMING	10	0.11	[0.00, 0.56]	12.4
QSO_nan	11	1.68	[0.01, 7.04]	2.64
QSO_BROADLINE	12	1.78	[0.03, 5.29]	6.54

Notes. Column 1: the name of the different subclass, constituted by the class name and subclass name. The subclass name 'nan' denotes classes with no specific subclass. Column 2: the label we used afterward. Column 3: the mean redshift of the subset. Column 4: the redshift range. Column 5: mean median signal-to-noise ratio, $\overline{SNR}$.

smaller than a few hundred. The classification and redshift pipeline of SDSS is based on a χ^2 minimization, by comparing each spectrum to the combination of basis templates, which are derived from rest-frame PCA of training samples (Bolton et al. 2012, B + 12 hereafter). The number of labelled spectra is more than four million.[1] In Table 1, we report the only 13 subclasses that have more than 20 000 classified objects, as this is the minimal sample size we need for the best training of our tools. Despite these representing a tiny fraction of the original class list (181), we stress that these 13 subclasses are representative of the most common objects one would expect to classify in typical spectroscopic surveys, especially if we look at the extragalactic sample. Most of the excluded classes, though, consist of stellar types (e.g. O, B star, dwarf, special carbon star, etc.) that have small observational samples collected, due to their intrinsic rarity. Of course, this is a limitation if one wants to apply the current classifier to real data that we expect to solve in the future by collecting more complete samples to build a compelling training sample, for example, using the early release of upcoming surveys (e.g. DESI and 4MOST). Also, the reduced number of subtypes adopted might not return the true final accuracy of the method, as we cannot predict if the classifier can perform closely to the average accuracy for all the missing classes. However, we believe that the number and variety of classes we have collected for this test, is already large enough to assess the potential of these (novel and unexplored) techniques. Indeed, since the main objective of this paper is to check if DL can efficiently and automatically classify spectra and measure redshifts of astronomical sources, the main conclusions we will draw will not be affected by the number of classes adopted, as long as the network can be trained for each class with a sufficiently large and representative knowledge base. Following this same line of argument, our results are also not affected by the accuracy of the classification performed in B + 12, as long as all spectra are assigned to a given class following self-consistent criteria. In this respect, GaSNet-II would just replicate the same classification bias intrinsic to the SDSS-DR16 sample, if any. However, from the perspective of the application to upcoming surveys, the problem of cross-contamination among classes needs

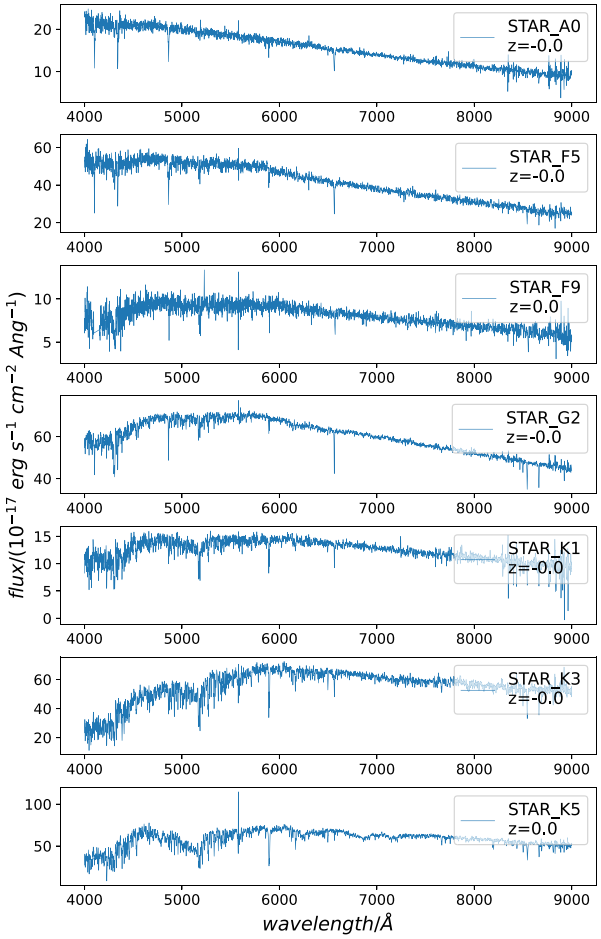

Figure 1. Example spectra of the seven stellar subclasses, corresponding to the first 7 of the 13 subclasses constituting the SDSS sample listed in Table 1. The A, F, G, and K stars with different subtypes are selected as the SDSS test samples to validate the ability of fine classification.

to be addressed to quantify how much this can impact the purity of classifications. Although this is not among the objectives of this paper, we briefly discuss this in Appendix A.

Finally, for the 13 suitable classes from SDSS-DR16, we can randomly select 20 000 spectra from each of these classes to collect a total catalogue of 260 000 spectra, constituting our primary data set. Most of the classes do not overlap physically, except for the 'BROADLINE' one, because if any galaxies or quasars have lines detected at the 10σ level with velocity dispersion (VDISP) σ >200 km s^{-1} at the 5σ level, the label 'BROADLINE' is added to their subclass.[2] The 20 000 spectra in each subclass are further split into random 70 per cent, 15 per cent, and 15 per cent subsamples to be used in training, validation, and testing, respectively.

Some typical spectra of different galactic (stars) and extragalactic (galaxies/AGN/QSOs) types are shown in Figs 1 and 2, respectively. The spectra are trimmed to stay within 4000–9000 Å wavelength range, for uniformity, then re-sampled to cover 5001 pixels, for a final effective binning of 1 Å pixel^{-1}. Besides this 'pre-processing' step, producing a uniform binned spectrum with respect to the original one, no additional data manipulation has been applied to the data. The range and mean redshift and SNR are listed in Table 1.

[1] DR16 Optical Spectra Overview.

[2] https://www.sdss3.org/dr9/spectro/catalogs.php

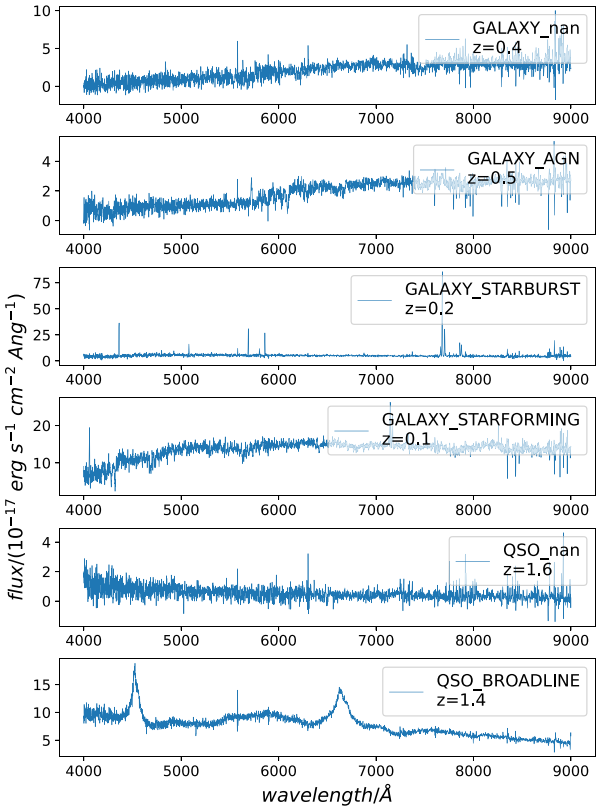

Figure 2. Example spectra of SDSS extragalactic subclasses, as listed in Table 1. We can clearly see the different features characterizing the different classes. From top to bottom, in particular, we can notice the increasing importance of the emission lines that play an important role in redshift prediction. The 'nan' type spectra generally lack such emission lines, although they might still contain some low-SNR ones, which are hard to see. This means that the 'nan' sample might overlap with other emission-line classes. QSOs also show a power-law continuum that does not carry any redshift information.

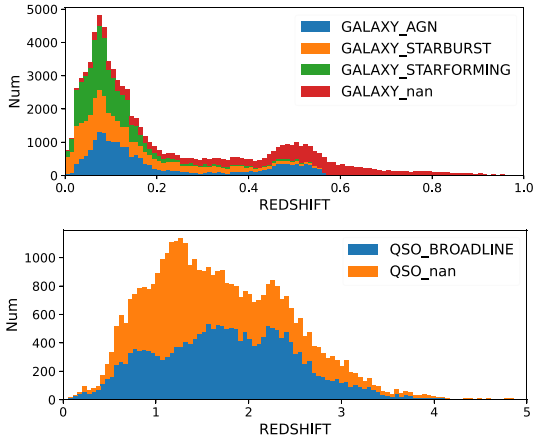

Figure 3. The redshift distribution of the SDSS-DR16 data set (stacked histogram). The mean and range of redshift are already shown in Table 1.

The distribution of redshift and SNR of different subclasses are shown in Figs 3 and 4. We stress that the high-redshift end on the redshift distribution in Fig. 3 is populated by a few systems. This is important to keep in mind, as we expect that this undersampling can

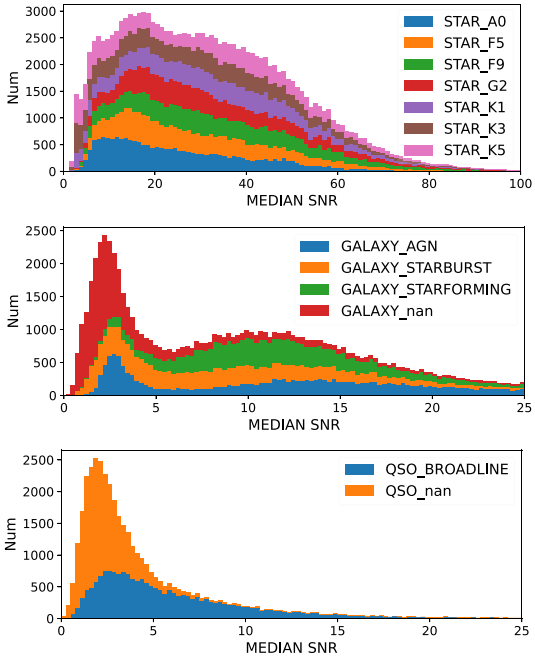

Figure 4. The SNR distribution of the SDSS-DR16 data set (stacked histogram). Top: star classes; middle: galaxy classes; and bottom: AGN classes. In general, the extragalactic objects are fainter than the star classes. The mean SNR is shown in Table 1.

impact the redshift predictions at the higher end of the class redshift distributions. On the other hand, the SNR distribution covers quite a high range, except for the QSO, which also shows a significant undersampling at SNR >10, and (counter-intuitively) causes worse predictions in this SNR range. Overall, to prevent such selection effects, one solution can be the use of simulated spectra, in order to collect a more balanced training data set. Although useful to solve these 'completeness' problems, this strategy has other limits which we will discuss in the next section, where we make use of 4MOST mock spectra, as an additional data set to test.

2.2 Other data set: 4MOST mock spectra

The data set consists of approximately 200 000 mock spectra obtained to reproduce 4MOST observation conditions, which are categorized into 10 different subclasses according to the adopted templates. We make use of a mock catalogue of spectra based on a customized software package,[3] reproducing the Exposure Time Calculator prediction of observed spectra for 4MOST. The software makes use of a series of customized templates selected for the different surveys (see Introduction) to be tested within the Extragalactic Pipeline working group (IGW8) and the Classification working group (IGW9) of the 4MOST consortium. The spectral wavelength range is cut to between 4000 and 9000 Å, and the number of pixels is interpolated to obtain 5001 pixels. The simulated spectra are generated from the given spectral energy distribution (SED) templates for a given set of observation conditions and random noise (including cosmic rays and randomized Lyα forest).[4]The spectral signal is obtained according to the exposure time and extinction: in particular, the exposure time is taken to be 1200 s for all spectra,

[3]https://escience.aip.de/readthedocs/OpSys/etc/master/index.html
[4]https://github.com/jkrogager/py4most

and the extinction is determined by the average galactic reddening law parametrized by Fitzpatrick & Massa (2007). The final sample contains a total of 10 subclasses, 5 galactic and 5 extragalactic. The galactic objects are: metal-poor stars and other dynamics tracers (Dyn) of The Milky Way Halo Low/High-Resolution Survey (Christlieb et al. 2019; Helmi et al. 2019), Cepheids in Magellanic Cloud (GalHR) of 1001MC Survey (Cioni et al. 2019), White Dwarf (ESN) of 1001MC Survey, Galactic disc stars (GalDiskLR) of 4MOST Surveys S1–S4 (Bensby et al. 2019; Chiappini et al. 2019; Christlieb et al. 2019; Helmi et al. 2019), and stars of Magellanic Cloud (MCsn) in 4MOST Survey S1 (Christlieb et al. 2019; Helmi et al. 2019). The extragalactic simulated sources are taken from mock catalogues and spectra provided by the 4MOST consortium extragalactic surveys: S5 eROSITA Galaxy Cluster Redshift Survey (Finoguenov et al. 2019), S6 AGNs (Merloni et al. 2019), S7 Wide-Area VISTA Extragalactic Survey (WAVES, Driver et al. 2019; Jin et al. 2024), S8 Cosmology Redshift Survey (Richard et al. 2019), and S10 The Time-Domain Extragalactic Survey (Swann et al. 2019). The respective contribution in simulated spectra of each survey is 2099 (24 658, 6056, 10 443, and 13 386) for S5 (S6, S7, S8, and S10). The templates used by S5, S6, and S8 were obtained by stacking spectra with the method from Comparat et al. (2020).[5] The stacked spectra were observed by SDSS within the Extended Baryon Oscillation Spectroscopic Survey (eBOSS) or the SPectroscopic IDentification of ERosita Sources (SPIDERS) programs (Almeida et al. 2023) and have similar properties to the selected targets to be observed by 4MOST consortium surveys S5, S6, and S8. As opposed to the SDSS-DR16, the classes available in the 4MOST sample are 'survey oriented'. In fact, the templates simulated come from different methods, and they are not purely grouped by physical properties, for example, star-forming versus passive galaxies or AGN, but rather customized for the survey requirements, including the SNR.[6] This is evident, for example, for the WAVES sample, which requires only redshift measurements of the targets, with the minimal exposure time and SNR needed to reach a reliable measurement. Table 2 shows the label of subclasses, SNR, and redshift distribution, while in Figs 5 and 6 we show some typical spectra from each of the 10 classes. The galactic objects have a higher average median SNR than the extragalactic objects. In the 4MOST sample, galactic objects exhibit a higher $\overline{\mathrm{SNR}}$ than those in the SDSS samples, whereas the extragalactic objects show a slightly lower $\overline{\mathrm{SNR}}$. The redshift distributions of the five extragalactic classes are shown in Fig. 7. The galaxy classes show a distribution that is similar to the one seen for the SDSS-DR16, while the quasars show a flatter distribution than the real data. As mentioned in Section 2.1, this might help alleviate the bias associated with incompleteness. However, this also raises the question of how realistic the 'prior' distribution adopted in simulation can be (e.g. see discussion in Li et al. 2022b, for imaging mock data). We postpone this test until we can access deep 4MOST observations, fully accounting for selection effects. Until then the 4MOST mock data set provides us a unique opportunity to test GaSNet-II as a general purpose 'survey-oriented' classifier, based on a large variety of classes, at the same time. Each subclass consists of approximately 20 000 spectra, which are split into 70 per cent/15 per cent/15 per cent for training, validation, and testing, respectively.

[5] https://github.com/JohanComparat/qmost_templates

[6] The main reason for this particular choice is that at the moment we have finished this work there was not yet a uniform physically motivated set of templates available for galactic/extragalactic targets in 4MOST, although a list of FGK star targets (from the galactic working group, IWG3) and a catalogue of stars with known labels for half a million stars from GALAH/APOGEE/RAVE/Gaia (from the ISSI team) will be available, and will be used for future GaSNet analyses. This does not represent a major issue for the purpose of this paper which aims to show the capabilities of the DL to perform classifications/regression tasks, regardless of the physics behind the spectra.

Table 2. 4MOST simulation data set.

Column 1	2	3	4	5
class_subclass	Label	$\bar{z}$	$[z_{\min}, z_{\max}]$	$\overline{\mathrm{SNR}}$
Dyn	0	–	–	74.5
GalHR	1	–	–	39.9
ESN	2	–	–	12.8
GalDiskLR	3	–	–	140.4
MCsn	4	–	–	72.3
COSMO_AGN	5	2.2	[0.9, 4.0]	6.3
ClusB	6	0.52	[0.3, 1.0]	5.8
WAVES	7	0.32	[0.0, 0.8]	1.6
RedGAL	8	0.33	[0.0, 1.1]	8.7
tides_host	9	0.11	[0.0, 0.6]	19.7

Notes. Column 1: the name of the different subclass. Column 2: the label we used afterward. Column 3: the mean redshift of the subset. Column 4: the redshift range. Column 5: the mean median signal-to-noise ratio, $\overline{\mathrm{SNR}}$. The first five subclasses are galactic objects and the last five are extragalactic objects.

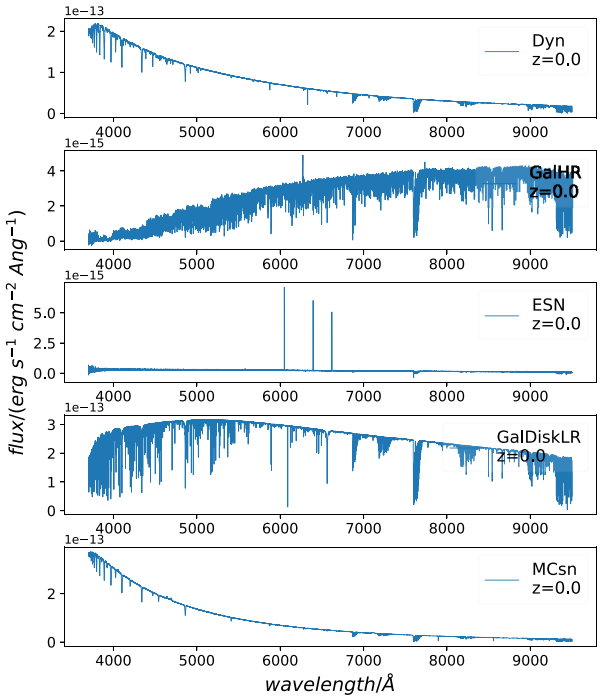

Figure 5. Example spectra of five galactic subclasses of the 4MOST sample, as listed in Table 2. From top to bottom, there are Dyn, GalHR, ESN, GalDiskLR, and MCsn.

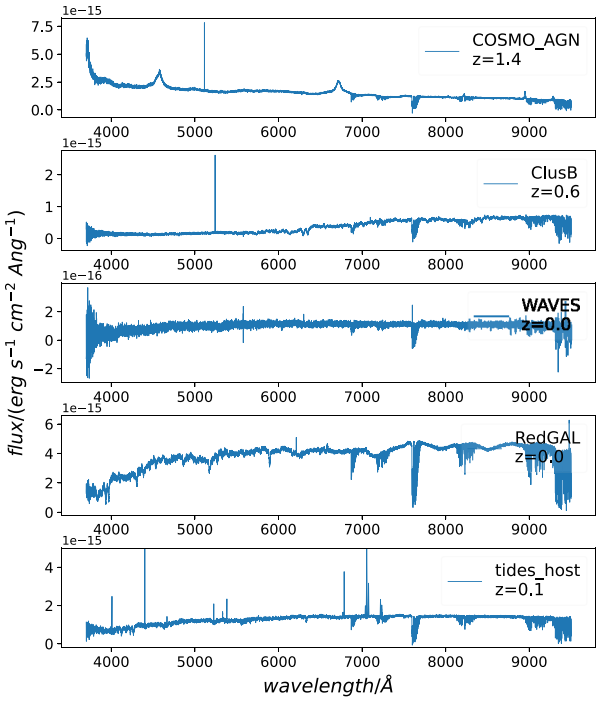

Figure 6. Example spectra of five extragalactic subclasses of the 4MOST sample (simulated), as listed in Table 2. From top to bottom, there are COSMO_AGN, ClusB, WAVES, RedGAL, and tides host.

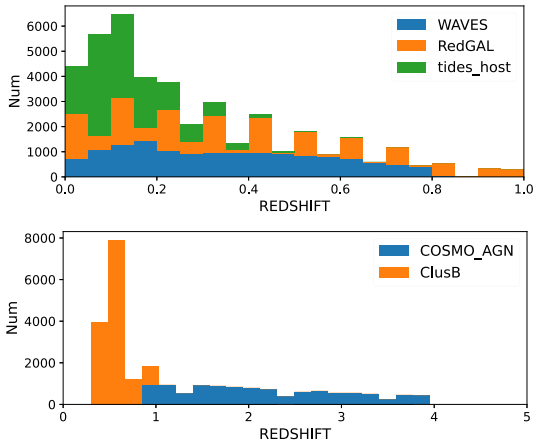

Figure 7. The redshift distribution of the 4MOST data set. The mean and range of redshift are already shown in Table 2.

2.3 Other data set: DESI spectra

The data set is constituted of 21 000 randomly selected DESI spectra,[7] which are categorized into three classes, QSO, STAR, and GALAXY. Each class consists of 7000 spectra in the data set. The DESI spectra are randomly selected from 'sv1' ('Target Selection Validation') samples and 'sv3' (One-Percent Survey) samples with SNR larger than 2 and *ZWARN* flat equal 0. Spectra are split into 70 per cent/15 per cent/15 per cent for training, validation, and testing, respectively. In the early data release version, DESI only provides a separation of the observed object into QSO–STAR–GALAXY, with only stars possessing further subclasses (eight in total), but with too few spectra to be used for training here. Hence, the DESI data set can be used to test GaSNet-II for a coarsely classified, poorly sampled data set (e.g. to be compared to a similar test on SDSS-DR16 as in Appendix B). The DESI classification and redshift prediction pipeline used REDROCK, a software package[8] based on fitting a set of PCA templates to every target at every redshift (DESI Collaboration 2023). The DESI spectra consist of three bands (*B*, *R*, and *Z* bands), with a wavelength range from 3600 to 9800 Å. Once again, spectra are interpolated to cover 5001 pixels in the wavelength range 4000–9000 Å, which are then used for the training. More details of the data set are shown in Table 3. The samples have a similar level of $\overline{\mathrm{SNR}}$ to the SDSS samples (Table 1) after the selection conditions were imposed. In Fig. 8, we show some spectra from the three different classes. Here, we have also highlighted, in different colours according to the legend, the subspectra collected from the three DESI arms, that are combined in the final DESI full-wavelength range spectra. Finally, the redshift distributions of galaxy and quasar samples are shown in Fig. 9.

Table 3. DESI data set.

Column 1	2	3	4	5
Class	Label	$\bar{z}$	$[z_{\min}, z_{\max}]$	$\overline{\mathrm{SNR}}$
STAR	0	–	–	19.1
QSO	1	1.59	[0.06, 4.27]	6.54
GALAXY	2	0.196	[0, 1.69]	7.53

Notes. Column 1: the name of different classes. Column 2: the label. Column 3: the mean redshift of the subset. Column 4: the redshift range. Column 5: mean signal-to-noise ratio, $\overline{\mathrm{SNR}}$.

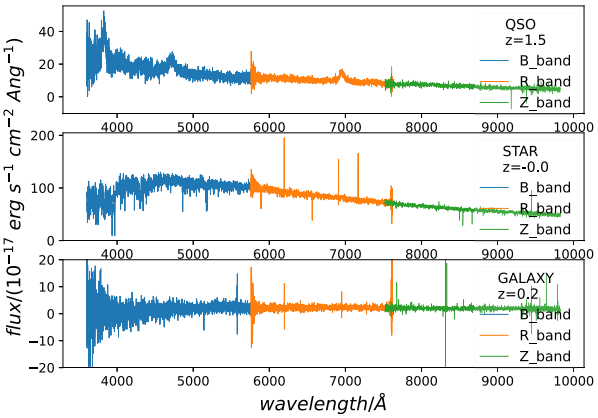

Figure 8. The typical DESI spectra of QSO, STAR, and GALAXY classes.

3 PIPELINE DESCRIPTION AND TRAINING

Thanks to their flexibility, efficiency, and accuracy, the multinetworks combination can be applied to the prediction of various astronomical parameters, and possibly form a fully automatic DL pipeline. The CNN (Krizhevsky, Sutskever & Hinton 2012) and the residual connection (ResNet; He et al. 2015) are two of the most widely tested DL architectures. CNN and ResNet have been extensively applied to classification and regression problems in astronomy, such as the photometric strong lens detection (Li et al. 2019, 2021; Petrillo et al.

[7] https://data.desi.lbl.gov/public/edr/

[8] https://github.com/desihub/redrock/releases/tag/0.15.4

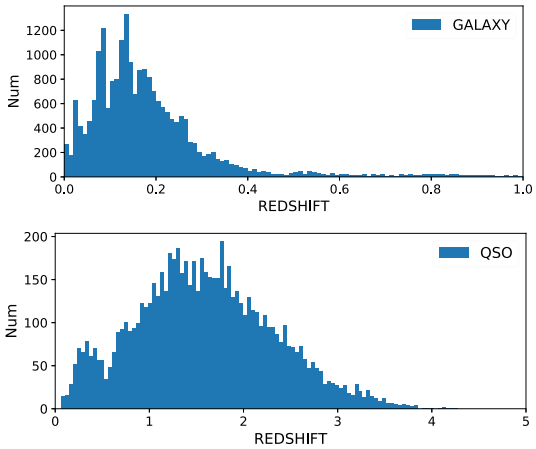

Figure 9. The redshift distribution of the DESI data set. The mean and range of redshift are shown in Table 3.

2019; Huang et al. 2020), galaxies morphology classification (Ball et al. 2004; de Diego et al. 2020; Domínguez Sánchez et al. 2022), star, galaxy, or quasars identification (Kim & Brunner 2017; Busca & Balland 2018; Parks et al. 2018; Guo & Martini 2019), photometric redshift predictions (Hoyle 2016; Pasquet et al. 2019; Li et al. 2022a), and stellar parametrization (Fabbro et al. 2018; Leung & Bovy 2019; Guiglion et al. 2024).

In this paper, we construct a multinetwork pipeline system, which is constituted by several, small, self-similar ResNet network models. The pipeline intends to map the pixel-level 1D spectra to return a classification probability and redshift. The classifier first is able to distinguish between subclasses. For instance, in the case of SDSS-DR16 (see Table 1), it separates the seven subclasses of stars (A0, F5, FG, K1, K3, and K5) that, being 'galactic' objects, are assumed to have redshift $z = 0$, and the 6 extragalactic objects, 4 of galaxies (nan, AGN, STARBURST, and STARFORMING) and 2 of QSOs (nan and BROADLINE). In total, there are 13 different classes. Then, on these extragalactic classes, GaSNet-II performs the redshift predictions and error estimates. Similarly, for 4MOST (see Table 2), the classifier separates the objects in the five star classes (Dyn, GalHR, ESN, GalDiskLR, and MCsn) and extragalactic classes (COSMO_AGN, ClusB, WAVES, RedGAL, and tides_host), then, for these latter, the GaSNet-II predicts the redshift and the errors. For DESI, the classifier just separates into three coarse classes (Table 3) and the redshift is measured for the galaxies and QSOs.

In this section, we introduce the details of the GaSNet-II architecture, the strategy for network training, and error estimates. We start by discussing in detail the training of the pipeline using the reference data set over which we want to test the capabilities of the pipeline, that is, the SDSS-DR16 sample. The structure and training of the pipeline will be the same for the other two data sets, that is, 4MOST and DESI, except that, due to the different numbers of labels (see Section 2), only the structure of the output will be different. For the latter data sets, we will discuss directly the performances on the test sample in Section 4.

3.1 GaSNet-II: philosophy and architecture

The philosophy behind the GaSNet-II architecture is based on two principles: simplicity and efficiency. Simplicity, because we want to build a network made of 'lighter', self-similar ResNets. The reason is that, by controlling each small network performance, we can easily check and control the whole pipeline performance. Also, having several ResNet blocks makes it easy to customize different subnetworks for different tasks. Efficiency, because GaSNet-II is able to parallelize classification and redshift predictions, which generally are part of a serial two-step process in classical pipelines, as the redshift accuracy is class dependent. Indeed, it is more difficult to determine the redshift for specific classes. An obvious example is passive versus active galaxies, as the former does not have as many high SNR features as the emission lines of the latter (Mateus et al. 2006).

To achieve this second objective, for GaSNet-II we decided to use a particular architecture made of parallel subnetworks, each one specialized on a specific task. This is sketched in Fig. 10(a), where a subnetwork is used to classify and give the probability to each object to belong to a series of pre-defined classes, while other parallel subnetworks, trained on each and only classes that need redshift estimates, are used to give the redshift predictions and error estimates. Obviously, the numbers of subnetworks are pre-assigned according to the number of those classes with redshift, that is, the training sample. In fact, being GaSNet-II a supervised network, the classes and redshifts need to be known as labels of the training sample used to train the networks.

However, all subnetworks are almost the same, in terms of their internal structure. Specifically, the multinetwork pipeline consists of one ResNet_P model to predict the probability, $\hat{P}$, of each subclass for classification, and six (identical) ResNet_i to predict the redshift, z, of different extragalactic objects, respectively. The index i corresponds to the label in Table 1. The input of all subnetworks are the 1D spectra, in flux units. As we will detail later, in this latter phase, GaSNet-II performs a Monte Carlo (MC) test, that allows us to estimate the errors, σ_z, on the redshift predictions. Hence, the output of the GaSNet-II pipeline is a 13-dimensional array of terns $(\hat{P},\ z,\ \sigma_z)$. The final input/output can be schematically summarized as:

$$F(\text{flux}) = \begin{cases} (P_i, 0), & i \in [0, 6], \\ (P_i, z_i, \sigma_{z,i}), & i \in [7, 12] \end{cases}, \tag{1}$$

where $\hat{P}_i$ are the probability from the ResNet_i classifier, z_i are the redshift predictions, and $\sigma_{z,i}$ are the redshift uncertainty, from the six ResNet regression models.

In terms of workflow, the classification is performed in parallel to (and hence independently from) the redshift prediction, hence this latter does not impact the classification. In principle, one can guess that this is a disadvantage as the knowledge of the redshift could improve the classification (for instance, this is easy to understand for stars that have $z \sim 0$). However, the GaSNet-II seems to reach already very high classification performances (~ 99 per cent, see Appendix B, Fig. B1) without this information. On the other hand, there are advantages of this 'parallel' approach: (i) one can scale-up the network by adding training samples for more classes, making it easy to extend the classification to other objects or even other targets, such as stellar parameters; (ii) parallelization reduces the impact of correlations between different quantities; (iii) for this reason, it is extremely flexible and can effectively applied to different SNRs and various surveys, as we will demonstrate later in this paper; (iv) it provides a reasonable uncertainty estimation, which is a robust starting point for subsequent Bayesian analyses; and (v) neural networks are powerful interpolators, thus also good at classifying spectra that lie within a learned multidimensional surface that cross-correlation would not grasp.

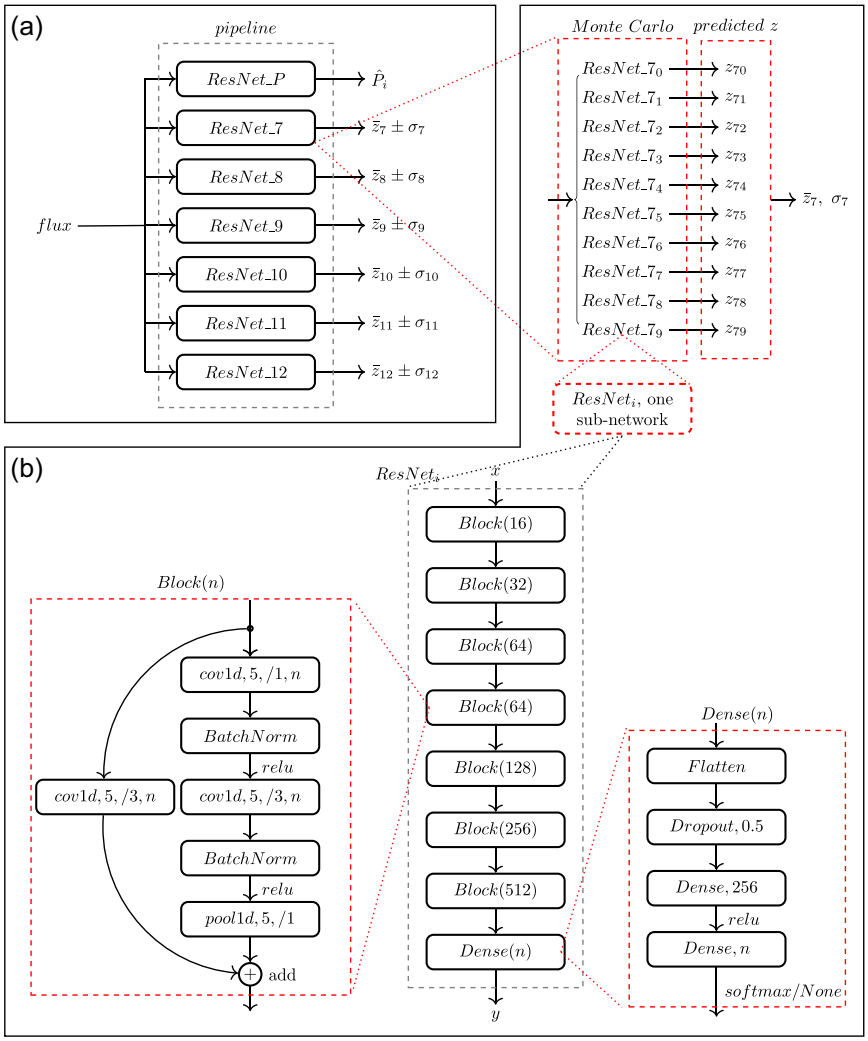

Figure 10. Panel (a): the general structure of the multinetworks pipeline. ResNet_P is used as a classifier and ResNet_$7-12$ is used for redshift prediction of extragalactic targets (note that ResNet_$0-6$ are missing because we do not need to predict the redshift of stars). One of the advantages of this structure is that it is simple and controllable, and can be trained and predicted in parallel. Panel (b): the detailed description of single subnetwork ResNet$_i$ (bottom figures) architecture, made by small blocks. The input of the network is 5001-pixel spectrum flux, and the output is the probability or redshift. The difference between classification ($n = 13$, softmax) and redshift prediction ($n = 1$, None) is the output dimension and the activation in the last layer. A feature-extract block Block(n) and a fully connected block Dense(n) are shown. cov1d is the 1D convolution layer. In one cov1d rectangle, 5 is the kernel size; /3 is the stride size; n is the number of channels. relu and softmax are the activate function, None represents no activate function here, that means liner. The left cov1d in the Block(n) shortcut is used to match the shape. pool1d is a 1D Maxpooling layer. As a schematic, the top right panel shows how to predict the redshift error of the label 7 (GALAXY_nan) subclass in parallel. Though 10 (customized) same subnetworks, trained by the same data but with different initial weights, 10 different redshifts were obtained from a single spectrum input. The expectation and error can be calculated. Other redshift errors are obtained in the same way.

3.2 GaSNet-II: pipeline description

In this section, we describe in detail the full end-to-end pipeline, which we have broadly described in the previous section. In the following, for brevity, we define the input of the subnetworks, x, and the fitting labels of subnetworks, y, as:

$$x = \mathrm{flux}/\sqrt{N}, \quad N = \sum_{j=1}^{5001} \mathrm{flux}_j^2, \tag{2}$$

$$\hat{y} = \mathrm{one-hot}(i), \quad i \in [0, 12], \tag{3}$$

$$y_i = z_i, \quad i \in [7, 12], \tag{4}$$

where the j represents the pixel index, from 1 to 5001, and the one-hot encoder converts the categorical data into digits, for example, one-hot(0) = 001, one-hot(1) = 010, one-hot(2) = 100, etc. In equation (2), the flux is normalized just like a vector. The fitting labels $\hat{y}$ are the labels converted by the one-hot encoder from Table 1. The fitting parameters y_i are the spectroscopic redshifts provided by the catalogue. To prevent the prediction of very high-redshift values, where the currently available training samples are too poor to give accurate results, we limit them to the range $z \in [0, 5]$. The loss functions used are

$$\mathrm{loss} = -\hat{y}_t \cdot \log(\hat{y}_p), \quad \text{(categorical cross-entropy)} \tag{5}$$

$$\mathrm{loss}_i = \begin{cases} \frac{1}{2}(y_{it} - y_{ip})^2, & |y_{it} - y_{ip}| \le \delta \\ \delta|y_{it} - y_{ip}| - \frac{1}{2}\delta^2, & |y_{it} - y_{ip}| > \delta \end{cases}, \quad \text{(Huber loss)}. \tag{6}$$

where $\hat{y}_p$, y_{ip} are the prediction values and $\hat{y}_t$, y_{tp} are the true values, and parameter $\delta = 0.1$. Huber loss combines the advantages of mean absolute error (MAE) and mean square error, and alleviates the sensitivity to outliers.

As seen in the previous section, the GaSNet-II pipeline is constituted by seven almost identical ResNet subnetworks. This is shown now in more detail in Fig. 10(b), where we offer a complete schematic view of the full architecture, which we describe below. Starting from the general structure seen in panel (a), we see that the subnetwork architecture consists of a series of ResNet 'blocks'. One of the advantages of using the subnetwork architecture, discussed in Section 3.1, is that it is particularly convenient to perform MC tests, which are the foundation of the GaSNet-II error estimates, as shown by the 'zoom-in' inset (top-right) in the same Fig. 10(b).

The idea behind the MC run is to use the different (10) subnetworks[9] with the same data, for example, a spectrum of an object of a given class, but with different initial network weights. In practice, the initial subnetwork parameters are set by a random Gaussian distribution, which establishes a random initial condition for the entire process, thus mimicking an MC experiment. However, this can also be seen as an ensemble training/MC, which is a relatively common practice in DL (e.g. Lakshminarayanan, Pritzel & Blundell 2016; Ganaie et al. 2021), and applied in the synthetic stellar spectra physical properties estimating(e.g. Bialek et al. 2020). This allows us to evaluate the stability of the output, by changing the initial condition of the training process. For the robust data points, different subnetworks are expected to predict values that are close to the ground truth, like the best-fitting values that find a global (or even a local) minimum in the χ^2 topology.

On the other hand, for the 'unstable' points different subnetworks are expected to find different predictions, like happens in best-fitting if the χ^2 has many local minima. In this way (despite the number of parallel experiments being only 10), we can separate the robust from unstable prediction targets. Hence, estimating the cumulative uncertainties on the final target estimates has two main objectives: (1) to associate a redshift and an error based on a probability distribution function (PDF) to every given target; and (2) to test the robustness of the network, by quantifying the overall predictions scatter with respect to the ground truth.

Indeed, from the 'zoom-in' inset of the MC test, in Fig. 10(b), we can see that the MC step provides a mean value, $\bar{z}$ and a variance, σ.

This is also done in parallel for the six extragalactic classes to obtain:

$$\bar{z}_i = \sum_{j=0}^{9} z_{ij}/10, \tag{7}$$

$$\sigma_i = \sqrt{\sum_{j=0}^{9} (z_{ij} - \bar{z}_i)^2/10}, \tag{8}$$

as shown in Fig. 10(b). The predicted expectations and errors will be shown in Section 4. In Table 4, we show the number of parameters, and the number of spectra adopted for the training of the subnetworks.

To check the effectiveness of the use of the mean redshifts and their errors from equations (7) and (8), we also provide the point estimate redshift for each target. These are based on a version of the network with only one ResNet in the MC module in Fig. 10(a) for each of the classes, and compare these with the ones we obtain from the MC run. These point estimates are analogous to individual measurements from standard techniques, like cross-correlation (REDMONSTER) or template fitting (REDROCK) and are meant to provide a realistic scatter of the estimates due to the combination of the data quality and the DL method.

Finally, Fig. 10(b) (top right) shows the convenience of the subnetwork architecture as the structure of each ResNet is the same for each one of the subnetworks, regardless of whether it is used to classify (e.g. ResNet_P) or to predict redshifts (ResNet_i $_{0-9}$).

[9]The choice of 10 networks is primarily to optimize the computational resources, to make GaSNet-II usable in small medium-scale servers with no much impact on the final results. For instance, considering the convergence of uncertainty in high SNR, Fig. 17 shows that 10 subnetworks are sufficient to robustly assess uncertainties and we do not expect to improve this result by increasing the number of subnetworks.

Table 4. Models detail. 'pars' is the number of network parameters. 'Num' is the number of training spectra used. 'loss' is the minimal loss on the validation set. 'acc' is the max accuracy on the validation set, and 'MAE' is the minimum mean absolute error on the validation set. [10]

Name	pars (10^6)	Num (10^3)	loss (10^{-3})	acc/MAE
ResNet_P	4.16	182	218	91.9 per cent (acc)
ResNet_7	4.16	14	0.868	0.011
ResNet_8	4.16	14	0.152	0.003
ResNet_9	4.16	14	0.066	0.001
ResNet_10	4.16	14	0.112	0.002
ResNet_11	4.16	14	10.2	0.107
ResNet_12	4.16	14	2.32	0.027

[10]Both trained by an NVIDIA Tesla P40 GPU

3.3 GaSNet-II training: SDSS-DR16

The training of GaSNet-II aims to minimize the loss function and maximize the accuracy (of the classification and predictions). As mentioned, all 'specialized' subnetworks are trained in parallel.

As a training set for the classifier network (ResNet_P, in Fig. 10a), we use a total of 182 000 SDSS-DR16 spectra, incorporating the 13 subclasses, each of them covered by 14 000 spectra for their training. By definition, each of the redshift prediction networks (ResNet_i in Fig. 10a), makes use of the same 14 000 used by the classifier for each subclass i, but with the purpose of mapping the input spectra to the labelled redshifts.

Under such partitioning of the training data, one can imagine that the classifier is set to search for the redshift in a larger parameter space, while the redshift 'regressor' networks, ResNet_i, are set to search for the specific redshifts of each subclass of spectra.

The result of the training process over the validation set is shown in Fig. 11, where a step learning rate is used. The learning rate starts at 10^{-3}, then slowly decays to 10^{-6} at the end (halving every 5 epochs when the epoch < 50) during 50 training epochs. The loss curves in the upper panel of the figure might indicate some slight overfitting, while the accuracy curves show that it does not affect the performance. The accuracy curve remains flat as more training epochs are implemented, meaning that it has achieved its upper limit. We have used a 0.5 dropout rate in the final layer to mitigate potential overfitting in the training set. Overfitting could be further reduced by using fewer network parameters or increasing the size of the training data (e.g. through online additive-noise data augmentation), however, due to the small amount of overfitting to correct we decided to test these strategies in future analyses. The checkpoints with maximum accuracy or minimum MAE are used as the model of the pipeline. Table 4 shows an average classification accuracy of 91.9 per cent

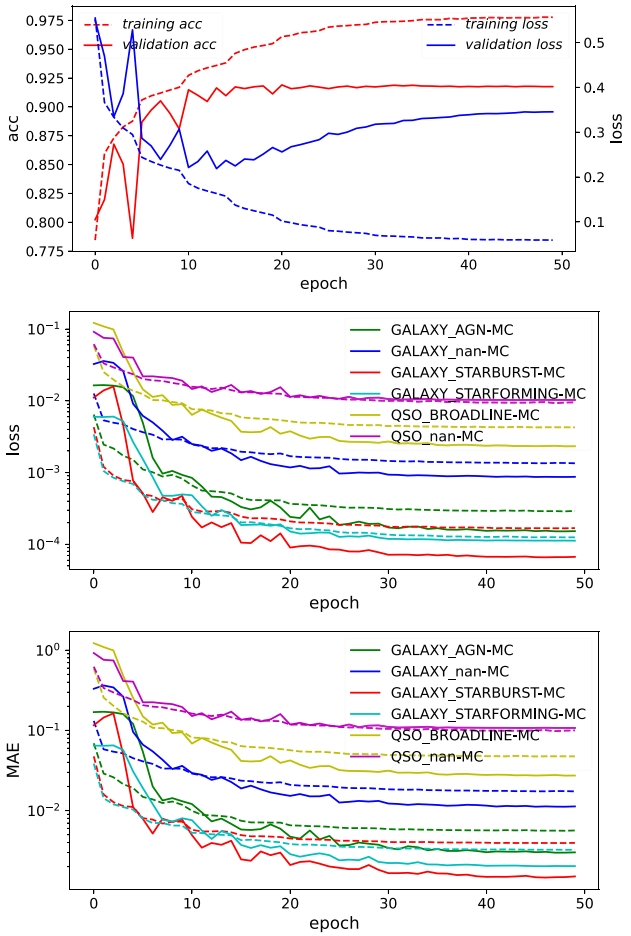

Figure 11. The training results for 50 epochs. We adopted a dropout rate of 0.5 in the dense layer to prevent overfitting during training. The first panel is the loss and accuracy of ResNet, which is used to classify the spectra. The second and third panels are the loss and the MAE of ResNet_i, which are used to predict the redshift. The dashed lines are the results of the training set, and the solid lines are the results of the validation set. The significant fluctuation in the first 20 epochs is due to the significant varying of learning rates. The overall worse performance in the training set is because we only employed the dropout in the training processes.

from ResNet_P, as well as a range of MAE for redshift estimation across different subclasses, ranging from 0.001 to 0.107. The number of trainable parameters of the subnetwork and the number of training samples are also provided.

4 RESULTS

In this section, we show the results of the pipeline using the same SDSS-DR16 data set and test sample. However, in the second part of the section, we also show the results of GaSNet-II, customized for the 4MOST mock data and DESI early data release, to demonstrate the potential for future application on Stage-IV surveys.

4.1 Statistical parameters

Before looking into the results, we introduce the statistical indicators to quantify the performance of GaSNet-II, specifically for the redshift accuracy. The first parameter is the Bias, defined as:

$$\mathrm{Bias} = |\ln\left(\frac{1+z_t}{1+z_p}\right)|, \tag{9}$$

where z_t represents the real value and z_p represents the prediction value. The Bias measures the deviation of z_p from z_t. In particular, we can use it to define the fraction of the 'good' estimates, Good_Frac (GF), as the fraction over the total number of spectra, N, of the redshift estimates for which the Bias is smaller than the related threshold, thr_x, that can differ for different classes (x = gal, qso). Hence,

$$\mathrm{Good\_Frac}\_x = \frac{N(\mathrm{Bias} < \mathrm{thr}\_x)}{N}. \tag{10}$$

We set the threshold of the galaxy species (nan, star-forming, starbursts, and AGN), thr_gal= 0.0015, such that optimal predictions are defined as Bias $<$ 0.0015, and the threshold of the QSO (nan and broad lines), thr_qso = 0.015, which qualify as good the predictions with Bias $<$ 0.015.

The second parameter is redshift relative bias Δz, defined as:

$$\Delta z = |z_p - z_t|/|1 + z_t|, \tag{11}$$

which is more intuitive than the Bias to interpret redshift discrepancies. In particular, this is closely related to the MAE, which is the mean of the Δz numerator, that is,

$$\mathrm{MAE} = \mathrm{Mean}(|z_p - z_t|). \tag{12}$$

As a reference, for the SDSS and DESI pipelines, $\Delta z < 0.01$ was essentially used as the catastrophic prediction threshold (Bolton et al. 2012; Dawson et al. 2016; Alexander et al. 2023), although it was less strict for high-VDISP QSOs.

4.2 SDSS-DR16 spectra

4.2.1 *Classification*

As discussed in Section 3, the ResNet_P subnetwork gives the classification probability (P_i) for each of the input spectra. This is the fastest task performed by GaSNet-II; it can perform the classification prediction of the 39 000 spectra belonging to the test sample in about one minute (excluding read time). The corresponding confusion matrix is shown in Fig. 12. Here we see that most of the subclass accuracies are larger than 90 per cent, except for the subclasses GALAXY_STARFORMING. The average accuracy of the 13 subclasses is 92.4 per cent. This average accuracy is certainly driven by the SNR of the spectra, as higher SNRs allow the network to better separate the spectra. This is shown in Appendix C, where we use the same GaSNet-II to classify increasingly higher SNR spectra and find that the average accuracy can reach a limit of $\sim$ 96 per cent for the highest SNRs. The star-forming galaxies are the class with lower accuracy, possibly due to a larger overlap (and thus a more uncertain classification) with other 'emission line' classes, e.g. AGN and starburst, but also with normal galaxies (GALAXY_nan class), possibly because low star-forming galaxies do not have strong enough emission lines to distinguish against non-star-forming systems. Some additional confusion can have a more physical origin, such as the smooth transition between AGN-dominated and host-galaxy-dominated signals. Furthermore, the accuracy of QSO_nan is also relatively low, but in this case, we track the reason for the typically low SNR, as seen in Fig. 4. Despite QSO_nan (GALAXY_STARFORMING) performing at 91 per cent (87 per cent) level, the missing sources are misclassified

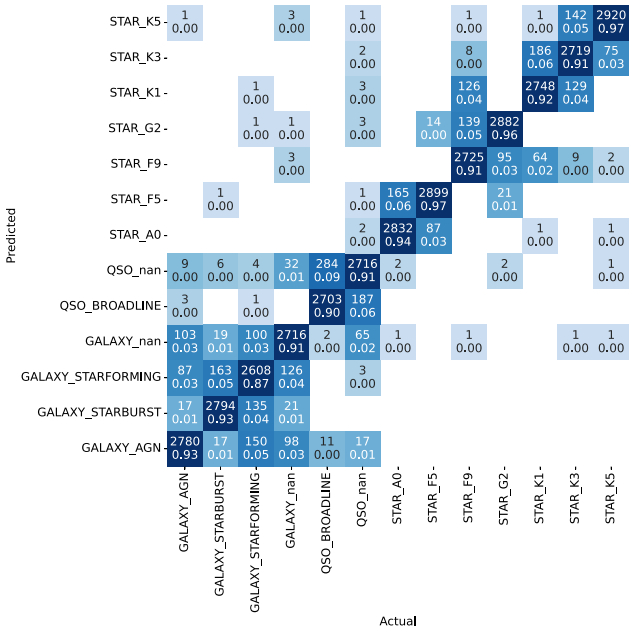

Figure 12. Confusion matrix results for the classification of the SDSS test set. The predicted and actual labels for each subclass (see Table 1) are listed on the left and bottom sides, respectively. Each subclass has 3000 test samples. The average accuracy is 92.4 per cent, and most are larger than 90 per cent (except the GALAXY_STARFORMING subclass). The matrix should be read along columns, that is the direction along which 100 per cent of the actual labels are distributed by the classifier.

as QSO_BROADLINE (GALAXY_nan, GALAXY_STARBURST, and GALAXY_AGN), which means that only the level of activity (intensity of the lines) moves some objects from one subclass to the other. If we also consider the arbitrariness in the separation of these subclasses in the SDSS classification, we believe that the accuracy reached by the GaSNet-II represents possibly a lower limit.

In Appendix B, we have collapsed all the subclasses on the three major classes of star/galaxy/QSO, which shows an average of 99 per cent accuracy. This test is important to reproduce the 'primary' coarse classification each of the forthcoming surveys will implement (see e.g. DESI in Section 4.4.1, for comparison). The main result is that a higher accuracy can be achieved (99 per cent on average) with fewer classes, using the same training data and network architecture.

4.2.2 *Redshifts: point estimates*

As anticipated in Section 3.2, we want to first derive the redshift point estimate for a single measurement from the spectra. This has an intrinsic error, which is due to a series of factors that we simplify into two categories: (1) SNR of the spectra and (2) measurement method. The former is linked to the structure of the data and how the features used for the redshift estimates are detected and measured (emission/absorption lines, 4000 Å break, etc.). The latter is linked to the accuracy of the method: for the DL tools, this lies in the impact of the weights and random seeds in the network. These two factors are not independent as, for instance, high SNR spectra make the impact of the weights minimal as the network tends to converge to a more robust estimate, and vice versa. Hence, the point estimate should reflect more the scatter due to these intrinsic sources of errors.

Fig. 13 shows the 'point estimate' redshift predictions of the six extragalactic SDSS-DR16 subclasses, described in Section 3.2. The overall impression is a rather good agreement between the predictions from GaSNet-II and the SDSS-DR16 redshifts, with rather small Δz and MAE, and a minimal fraction of catastrophic estimates, except for the QSO_nan subclass. The best accuracy is found for the GALAXY_STARBUSRST and GALAXY_STARFORMING, subclasses, $\Delta z = 0.001$, while QSO_nan shows the worst $\Delta z = 0.047$. These accuracies are still about one order of magnitude larger than the ones required for redshift catalogues (see e.g. Bolton et al. 2012), but this is not a major concern for this analysis that is not meant to optimize the redshift accuracies.[11] The GF is generally larger than ~ 50 per cent but reaches 80 per cent relevant fractions only for three classes. We see an increasing scatter of the predictions at higher redshifts in almost all categories, mainly driven by the poor coverage from the training samples of high redshifts. As we will see, training on mock spectra can strongly alleviate this problem. The relatively poor performance of the QSO_nan sample, as we mentioned above, is additionally driven by the low SNR of the spectra. As we will discuss later, the SNR has a large impact on the accuracy of the predictions.

The values of Bias of all subclasses are shown in Fig. 14. In this figure, we present the Bias values as a function of the redshift and colour-coded by their SNR. The GF is reported in the legend for each SNR bin. It is evident that the number of 'good' predictions increases with SNR, which also correlates with redshifts; the lower SNR spectra generally correspond to the higher redshift ones. This also explains why even classes with lower GF, like the GALAXY_nan (GF = 0.63), reach a rather large GF$\sim$ 90 per cent, for SNR > 10 spectra. If we exclude the QSO_nan, which has too few SNR > 10 spectra to have reliable statistics (see Section 2.1), all classes have GF going between 63 per cent and 94 per cent, while the average GF is larger than 90 per cent for starburst, star-forming, and broad-line QSO, clearly because of their well detectable emission lines. On the other hand, the lower accuracy of the normal galaxies (GALXY_nan) is due to the fact that GaSNet-II learns the redshift mainly from the continuum shape and possibly the absorption lines, whereby the spectra have lower SNR for key features compared to the emission-line galaxies; this can limit the performance of the former subclass.

4.2.3 *Redshifts: MC estimates*

We finally discuss the redshifts and errors of the six extragalactic subclasses predicted by the MC test discussed in Section 3.2, which are shown in Fig. 15. The main evidence emerging from a quick view of the predicted values is that the accuracy is comparable to the point estimates, as measured by MAE and Δz, which are very close, or even identical to the ones shown in Fig. 13. Looking at the errors, they are extremely small for the predicted values that distribute along the 1-to-1 relation and become bigger for the (few) highly scattered predictions.

As discussed in the previous section, QSO_nan is the most problematic subclass, showing a larger scatter, and, consequently, larger errors. Looking at the high-z end in all classes, we see the effect again of the sparse training samples which contribute to the larger errors, which are mirrored by the increased scatter in the estimates already noticed in Section 4.2.2. This is quantified in Fig. 15, where the upper panels show the mean σ_z of the redshift estimates in different redshift bins. Here, we can clearly see that the mean errors increase with increasing redshift in almost all classes, except the GALAXY_AGN. Some points' errors are underestimated,

[11] Preliminary tests considering anomaly detection show that we can achieve $\Delta z \sim 10^{-4}$. This will be discussed in upcoming analyses.

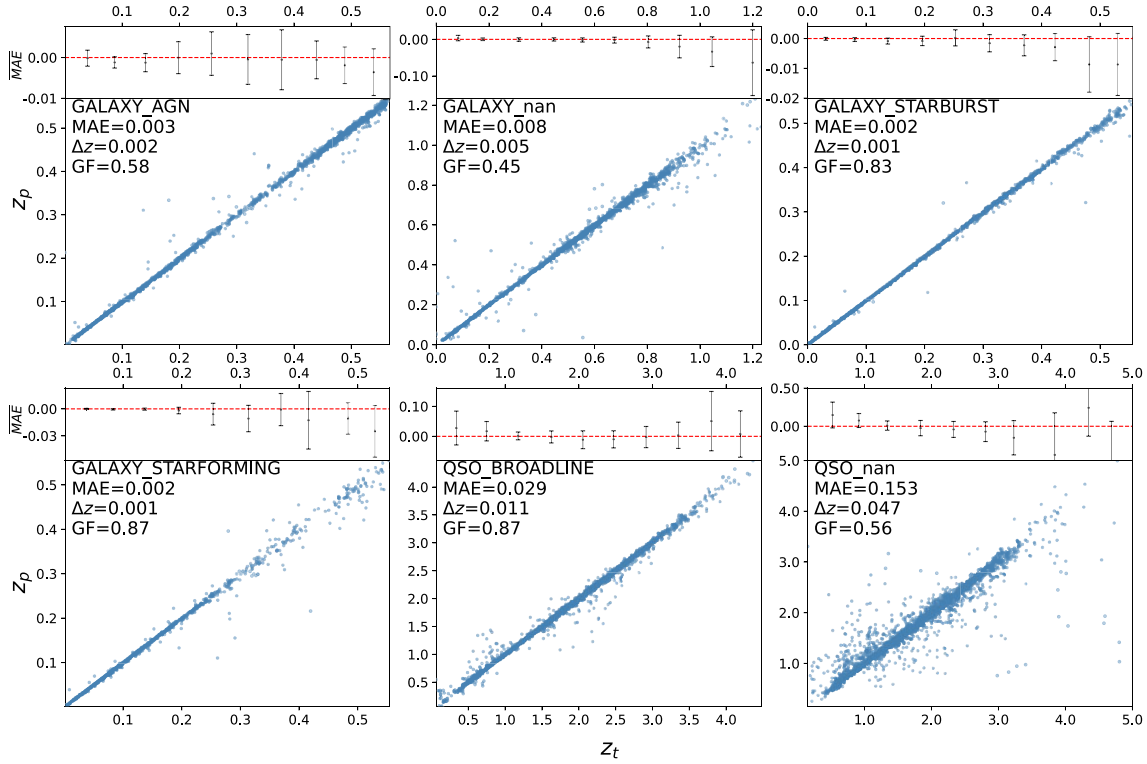

Figure 13. Redshift predictions of six extragalactic SDSS subclasses, each of which used one subnetwork. The subclasses GALAXY_STARBUSRST and GALAXY_STARFORMING have the best redshift estimations, with an error of $\Delta z = 0.001$. This can be attributed to the presence of significant emission lines in their spectra, as shown in Fig. 2. The subclass QSO_nan has the worst estimation with an error of $\Delta z = 0.047$. This subclass is characterized by the lowest SNR, a high-redshift range (Table 1), and a weaker broad emission-line signal in the spectrum (Fig. 2). Error bars on each redshift bin (10 bins) are plotted at the top of the panel. The MAE for each bin is used as the error bar. The plot clearly indicates that errors become significant at the higher redshift end, which is attributed to the lack of training samples in that region.

particularly at the high-redshift end. This is due to a lack of training samples in those regions, which results in lower accuracy in this region. The bottom line is that the estimated errors are indeed a measure of the reliability of the GaSNet-II predictions, as large error bars emerge either because the estimated values are far from their true value, or because the predicted value is poor due to the poor knowledge base. In particular, we can use the estimated error, σ_z, to determine whether an estimate is 'robust' or 'unstable', using the MAE (listed in Table 4) as a lower limit for an estimate to be unstable.

Before we discuss the predicted errors, we want to see whether the mean redshift estimates behave similarly to the point estimates, or, in other words, whether the point estimates are drawn by the redshift PDF derived by the MC run. This is needed to check if the point estimates are 'unbiased' predictions of the 'ground truth'. To do that, in Fig. 16, we plot the relative scatter normalized to the errors, $t = |z_t - \overline{z}_p|/\sigma_z$, for the different test sets, which should be enclosed in the range [0, 3] for a Gaussian distribution. Here, we see that the great majority of the point estimates are within the $3\sigma_z$ distribution with fractions of the order of 0.96 or higher. This is not fully compatible with a pure Gaussian distribution (expected to be ∼0.99), but rather shows some excess outliers, which we can roughly estimate to be no more than 5 per cent. Also, we see that some subclasses are more prone to systematics than others, like the 'GALAXY_AGN' and 'GALAXY_STARBURST', that have a tendency to provide overestimated 'point' redshifts. We stress here that the point estimates are obtained by a separate, independent pipeline, trained to optimize the redshift estimate on a single run, so they cannot be considered a random sample of the MC run, which is trained to optimize the mean $\overline{z}$. We take this into consideration in the discussion below.

Moving to the error estimate, we start by connecting these errors with the data structure. If the errors are artificially produced by internal network errors, due to the stochasticity of some processes, then these should not have any correlation with the spectra uncertainties. To show that, in Fig. 17, we compare the σ_z and SNR of the spectra, where we see a correlation between the error size and the SNR, as quantified by the median values (dashed line), showing that the lower the SNR the larger the σ_z tends to be. This is proof that the errors are driven by the data noise, which was assumed without proof so far in this section, and is consistent with the impact of the SNR in classification, discussed in Section 4.2.1 and Appendix C. However, at any fixed SNR value, we also see the scatter of the σ_z from class to class, with the QSO generally showing larger errors. If we exclude the regions with sparse sampling (see e.g. SNR ∼ 5 for 'GALAXY_STARFORMING', or SNR∼6–8 for 'GALAXY_AGN'), where the larger scatter of the errors might reflect lower precisions due to a poor training sample, the reason of the σ_z variation from class to class should reside in the type of features that GaSNet-II used for the predictions. For instance, in the case of the 'QSO_BROADLINE' (and perhaps also partially true for 'QSO_nan') it is the line broadening that leads to more insecure estimates, especially at lower SNR. Interestingly this is not seen for 'GALAXY_nan', which lets us speculate that for these

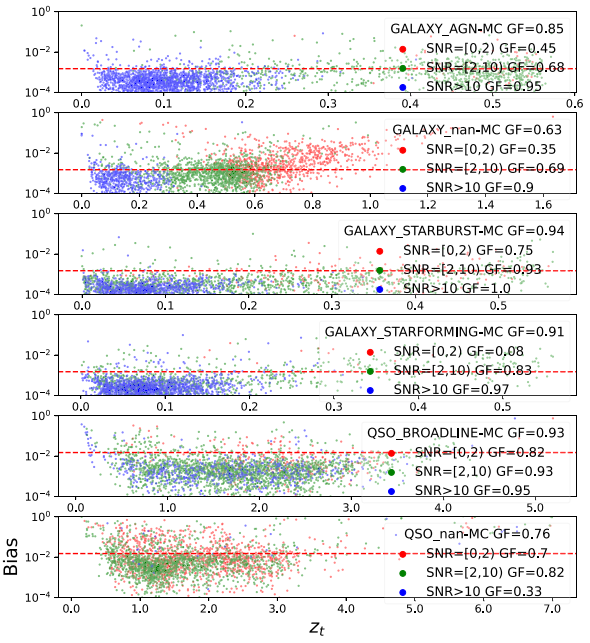

Figure 14. Bias as a function of the redshift for different extragalactic SDSS classes as indicated by the legend on the right. The spectra are divided into low (SNR=[0,2)), medium (SNR=[2,10)), and high (SNR>10) SNR to show the performance at different noise levels. The GF within each SNR bin is reported in the legend. The plot shows clearly that estimate deviations exhibit more scatter as the SNR decreases, implying larger statistical errors. The errors increase at the high-redshift end, where the SNR is typically lower. Another source of scatter is that as the redshift increases, the training samples become smaller. See also Section 2.1.

systems the absorption lines are not driving the redshift estimates, but rather the full spectrum and there is a smooth and regular degradation of the errors for smaller and smaller SNRs, similar to what is seen for GALAXY_AGN. Direct analysis of the impact of the spectral features on the accuracy is beyond the purpose of this paper and would require more sophisticated techniques like self-attention methods of anomaly detection, which we will address in forthcoming analyses. However, to give a preliminary insight into the importance of the spectral features in classification and redshift predictions, in Appendix D, we show the gradients of the classification probability and the output redshift with respect to input flux, which allows us to visualize the impact of spectral features in the GaSNet predictions, although they cannot give a real measure of the impact of the continuum.

On the other hand, GALAXY_STARFORMING seems to be insensitive to SNR until they reach SNR ~ 7, below which prominent emission lines start to blend into the noise, and then the continuum takes over dominating the larger errors, similar to GALAXY_nan. We also notice different behaviour between GALAXY_STARFORMING and STARBURST, as, for the latter, σ_z is increasingly noisier toward low SNR. As the most important features for these two classes are the emission lines, one would expect a similar behaviour for σ_z. There are two reasons for this: (1) emission lines in starburst galaxies dominate the spectra and GaSNet-II does not learn much from the continuum for star-forming systems. Thus the redshifts are fully determined by the ability of the ResNets to cross-correlate emission lines over a large wavelength range; and (2) ResNets is perhaps not the ideal tool for this emission-line redshift estimation task, which is typically well handled by other DL structures, like 'self-attention' networks (e.g. Han et al. 2020). We will discuss this in detail in Section 5. Finally, another source of uncertainty in both redshift and classification can be the VDISP, as this can produce a different broadening of the line that might reduce the accuracy of both tasks. In Appendix E, we demonstrate that both σ_z and classification accuracy show almost no correlation with the VDISP, inside the different classes.

The bottom line is that the estimated error sizes as a function of SNR and redshift seem to be mainly driven by the data quality and data features as one should expect from standard analysis methods, rather than the stochasticity of the DL network. As a consequence, we are motivated to use σ_z as a proxy of the 'robustness' of the redshift estimates, as we now can interpret σ_z as the cumulative effect of the variance of the weights of the network (see Section 3.2) and the data noise. Also, we can expect that the estimates with smaller σ_z are more tightly distributed around the true value. In Fig. 18, we show again the Bias versus z, which is split into 'robust' or 'unstable' categories based on whether their $\sigma_z \leq$ MAE or $>$ MAE, respectively, where MAE is the mean absolute error in the validation set (Table 4). The robust limit is very close to the GF limit, and only in the 'GALAXY_nan' or 'QSO_nan' subclasses it is significantly larger. Thus, the robust estimates have a fraction over the total samples that are larger than the GF defined by the Bias threshold. This result is particularly relevant for practical applications, as for new spectra with no *a priori* information on the redshift, the use of the redshift errors proposed here allows us to discard unstable estimates (larger deviation points) without knowing the ground truth.

4.3 4MOST mock spectra

Next, we analyse the 4MOST data set introduced in Section 2.2. The main reason to use this data set is to test GaSNet-II with spectra close to expected data from major Stage-IV upcoming spectroscopic surveys, but classified on the basis of the survey requirements, thus providing a different classification approach, more survey-oriented. Overall, this would allow us to test the versatility of the pipeline, to respond to different requirements, both in classification and in redshift predictions.

The training of GaSNet-II with the 4MOST spectra follows the same procedure discussed for SDSS-DR16 in Section 3.3. As anticipated, the size of the sample for each class (total of 10 classes) is the same as SDSS-DR16 (20 000) and we use the same training, validation, and test division (70 per cent, 15 per cent, and 15 per cent).

In the 4MOST observation phase, the labelled training data rely on the classification of the first months of 4MOST observations to develop a customized training sample based on data collected from the different survey teams. Alternative approaches might rely on the use of mock data, or using visually classified data.

4.3.1 Classification

Starting with the classification, in Fig. 19 we show the confusion matrix obtained over the test samples. GaSNet-II achieves an accuracy beyond 90.0 per cent for the majority of subclasses, and an average overall accuracy of 93.4 per cent, which is slightly better than the one found for SDSS-DR16 (92.4 per cent). One reason can be the absence of contamination discussed above, which we will address at the end of this section; another reason is likely to be the even stronger disparity in SNR between subclasses. Before we check that, we first discuss some other relevant features from the confusion matrix.

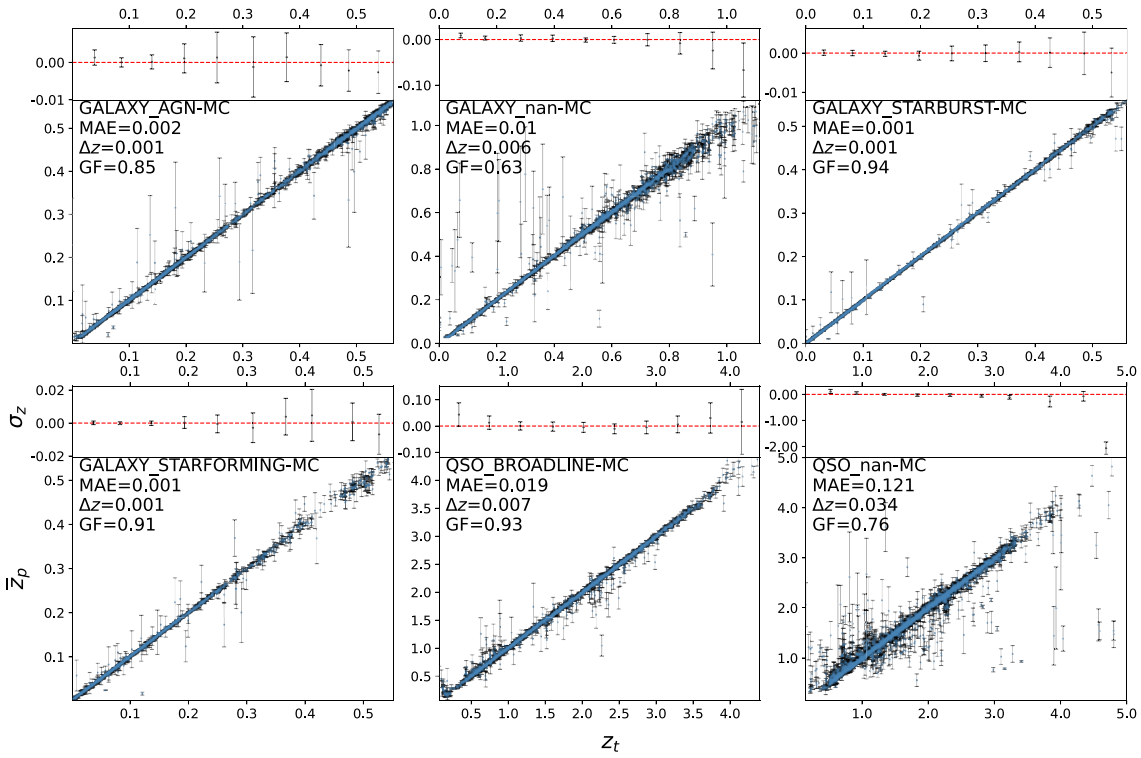

Figure 15. The mean redshift predictions and errors of the six extragalactic SDSS subclasses. The error bar of each sample point represents the standard deviation obtained from the MC estimation of 10 subnetworks. In the top left of each main panel, the subclass name, MAE, Δz, and the GF are displayed. The points in the top panels display the mean of the distribution of the $\bar{z}_p$ residuals ($\bar{z}_p - z_t$) with respect to the true values (z_t) in each bin, and error bars corresponding mean σ_z values (see the text).

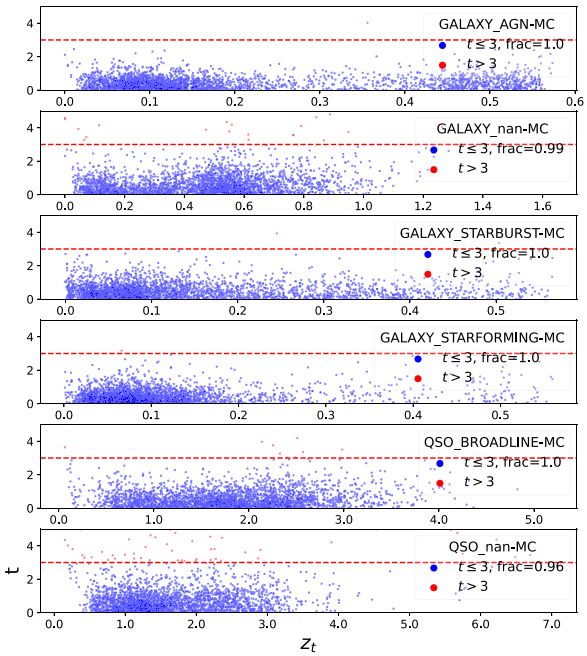

Figure 16. The distribution of t versus redshift, where t is defined as $t = |z_t - \bar{z}_p|/\sigma_z$. In the legend, 'frac' denotes the proportion of the sample with $t \leq 3$.

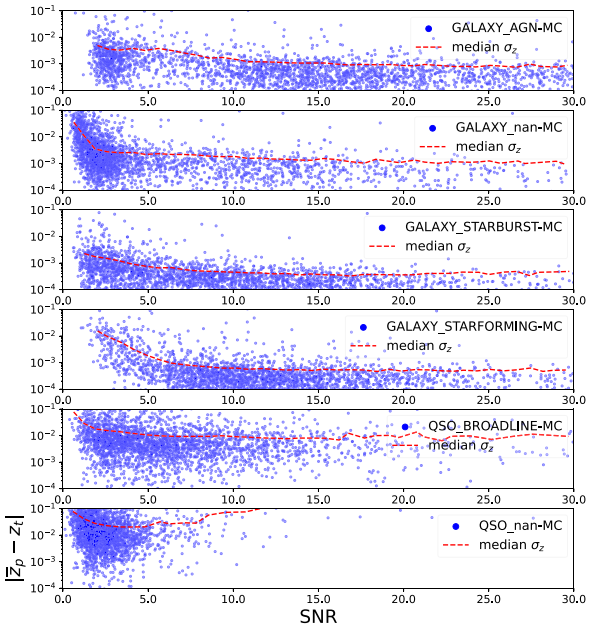

Figure 17. The distribution of $|\bar{z}_p - z_t|$ versus SNR for the SDSS test data, tracking the performance of error estimations in different noise levels. Median σ_z is indicated by a dashed line. It demonstrates the MC method can reflect the uncertainty realistically. In low SNR regions, the value of median σ_z is larger compared to the high SNR regions, as expected.

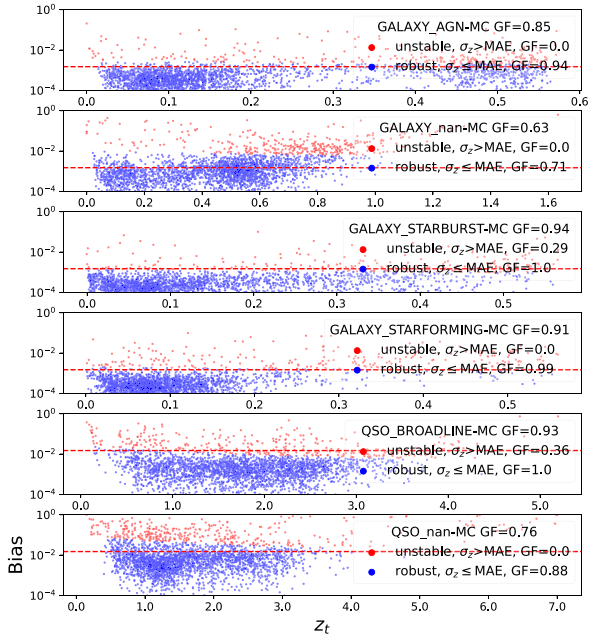

Figure 18. Bias of the 10 subnetworks used. The *x*-axis is the real redshift and the *y*-axis is the Bias. The MAE is listed in Table 4. 'robust' is defined as $\sigma_z \leq$ MAE, where MAE is the mean absolute error in the validation set. This demonstrates that unstable points (larger deviation points) can be automatically found without knowing the ground truth.

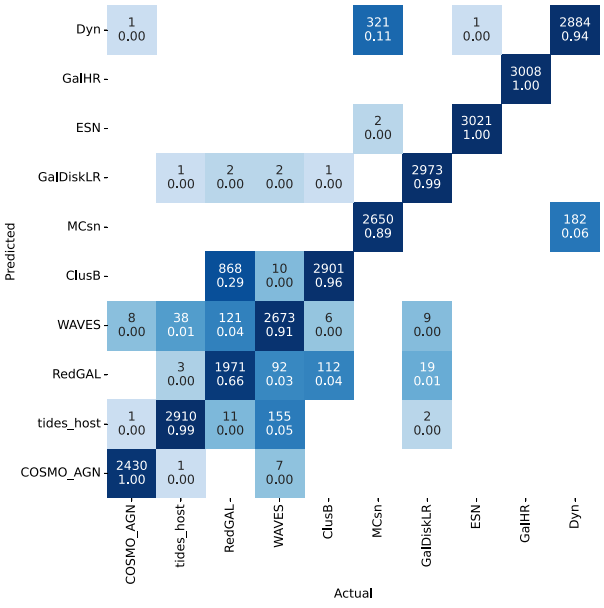

Figure 19. The figure displays the classification results of the 4MOST model on the testing set. It presents a confusion matrix where the legends are the same as Fig. 12. This figure indicates an average accuracy of 93.4 per cent. The worst performance is observed in the subclass RedGAL, which has an accuracy of only 66 per cent. 29 per cent of the spectra in RedGAL are misclassified as ClusB. Note that the matrix has to be read along columns, that is the direction along which the 100 per cent of the true labels are distributed by the classifier.

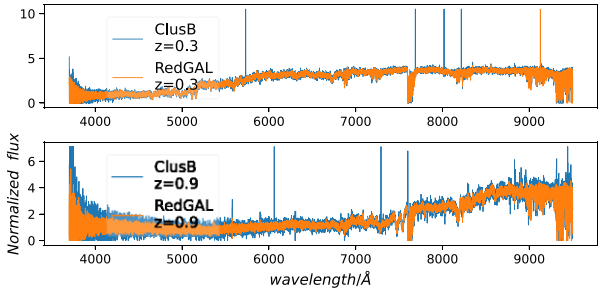

Figure 20. We randomly pick four spectra. The upper panel shows ClusB and RedGAL spectra with a redshift of 0.3. The bottom panel shows the spectra with a redshift of 0.9.

In particular, we notice a striking 100 per cent score by the COSMO_AGN class that is superior to the 91 per cent scored by the GaSNet-II on the SDSS AGN sample. Since the mean SNR of the two data sets is very close in galaxy and AGN, (see Tables 1 and 2), we identify the reason for this overperformance on the COSMOS_AGN sample to the different redshift distributions, whereby the 4MOST sample lies at a much higher average redshift compared to the SDSS AGN. This makes it easier for GaSNet-II to unequivocally distinguish the brightest AGN features from, for example, starburst/star-forming galaxy emission lines, for faraway systems than for closer ones. However, another factor that might help this outperformance is the limited chance of cross-contamination among the training/testing classes, which have been constructed here on distinct templates to obtain the mock spectra (see also below).

The only clear case of such contamination is the mixing between subclass 'ClusB' (label 6, corresponding to bright cluster galaxies) and 'RedGAL' (label 8, i.e. red galaxies). ClusB likely systems are a peculiar subsample of the RedGAL systems, at least at low redshift, as bright central cluster galaxies are generally old, red galaxies, particularly in their centres (see e.g. Bernardi et al. 2007), which is where 4MOST fibres would be placed. Fig. 20 shows the templates of two ClusB spectra at redshift 0.3 and 0.9, respectively versus two redGAL templates at the same redshifts, normalized to the same flux at 6000 Å at each redshift. We are asking the classifier to separate spectra that are nearly indistinguishable at the same redshift. Surprisingly, in Fig. 19, we see that GaSNet-II can correctly predict the clusB galaxies, while it confuses the RedGAL for ClusB in $\sim$ 29 per cent of the cases. We can possibly explain this with the fact that ClusB galaxies often systematically show emission lines in their spectra, while the RedGAL mostly do not (see again Fig. 20), hence we argue that the emission lines are features that GaSNet-II associates to ClusB galaxies and not RedGAL, where they are not dominant. This means that RedGAL spectra with emission lines have a larger chance of being classified as ClusB. To conclude this section, we refer the reader to Appendix B where, as for SDSS, we have performed the classification of the spectra by grouping the different star, galaxy, and AGN classes to emulate a coarse STAR-GALAXY-AGN classification, to be compared with a similar one from SDSS and DESI. We stress here that this experiment, besides putting the performances on 4MOST templates in the context of other reference surveys, provides us also a test on a more physically oriented sample, rather than a survey-oriented classification discussed so far. This is closer to what GaSNet will be required to perform in the early stage of 4MOST operations. In this case, we can see that the coarse classifier can reach an even higher mean accuracy of 98 per cent, comparable with what we have seen for SDSS.

4.3.2 Redshifts

We finally show the results for the redshift predictions, limiting ourselves to the MC estimates with errors. In Fig. 21, we show the predicted redshifts for all the 4MOST extragalactic subclasses. The figure indicates an average Δz of 0.0055 for galaxy types (ClusB, WAVES, RedGAL, and tides_host), while it becomes 0.003 for AGN. The average GF for the galaxy is 0.68, while for AGN is 0.71. These latter are the class for which GaSNet-II also provides the most accurate classification, meaning that the combination of good SNR and emission lines, permits high performances for both tasks. Among the galaxy types the average error is dominated by the WAVES class which has the largest errors, possibly due to the low average SNR (see Table 2). The same WAVES class also shows the highest relative scatter $\Delta_z = 0.014$ compared to $\Delta_z \sim 0.004$ shown by the majority of the other subclasses. Overall the Δ_z found for the 4MOST mock sample seems slightly worse than the one measured for SDSS ($\Delta_z \sim 0.003$), although a direct comparison is not appropriate, with the two samples having different observational constraints, especially in terms of SNR, for instance, 4MOST AGNs and galaxies have a lower SNR except the 'tide_host' subclass (e.g. comparing Tables 1 and 2). The 4MOST redshifts also show a GF on average slightly lower than the one of SDSS as reported by the mean good fractions in the legends of Fig. 21, against the GFs reported in Fig. 15, for SDSS. Once again the WAVES spectra are the ones with the worst GF, which are a consequence of the systematically larger errors, ultimately driven by the low SNR.

As for the comparison with standard methods, here a full detailed check of the relative performance of GaSNet-II with respect to tools like REDROCK and REDMONSTER is beyond the scope of this paper. However, to put the GaSNet-II performances into perspective, on a series of benchmarking tests on simulated 4MOST consortium data sets, we have found GaSNet-II GF to be ∼20 per cent worse than REDROCK and REDMONSTER, although, for some classes, like AGN/QSO, GaSNet-II shows a GF even better than classical tools. For instance, REDMONSTER shows an average GF of 0.71 (GF for AGN/QSO is 0.43), mean absolute deviation (MAD) of 0.00042, and Time (in the unit of seconds per spectrum per core, sec/spec/core) of 1.02; REDROCK shows an average GF of 0.48 (GF for AGN/QSO is 0.23), MAD of 0.051, and untested Time; while for GaSNet-II we find an average GF of 0.40 (GF for AGN/QSO is 0.70), MAD of 0.0086, and Time of 0.00089 on the AGN/QSO/GALAXY redshift test sets. This indicates that there is still room for GaSNet improvements, which can be consolidated with final, more sophisticated, mock data, and eventually with the first 4MOST observations.

4.4 DESI spectra

We finally apply GaSNet-II to the early release DESI data. As seen in Section 2.3, the DESI classification taxonomy is less complex only a very broad classification (i.e. star, galaxy, and quasars), and their numbers are less abundant, as we could test our tools over ∼1050 classified spectra for each class. This allows us, besides testing GaSNet-II on a further data set, with a different observation set-up and size, to perform a basic analysis over a 'coarse' classification which is similar to what we expect to implement for 4MOST earlier data releases (see also Appendix B). The classification and redshift estimates are quickly discussed below.

4.4.1 Classification

The separation of the test sample on the three DESI classes is shown in Fig. 22, where the confusion matrix indicates the accuracy of each of the three classes is larger than 93 per cent, and the average accuracy is 96 per cent. The high accuracy is obviously highly dominated by the small number of classes, however, this also shows an almost absent ambiguity of the classification for classes notoriously prone to confusion, for example, stars and galaxies. This is likely due to the ability of GaSNet-II to guess the redshift and (eventually) the shapes of the spectral features. We expect though that with a larger training sample the accuracy will be further increased. To put these results in perspective with other data sets, in Appendix B we have performed a similar analysis for SDSS-DR16, by collapsing all spectra subclasses into three broad classes as for the DESI data set. We anticipate that, using the same number of SDSS training samples, we find a 99 per cent accuracy for such a coarse classification, that seems rather higher than the one obtained for DESI. This implies that the quality of the spectra, rather than the number of training samples, is the major factor contributing to the accuracy. We expect to return to such a test in upcoming DESI releases to confirm this result.

4.4.2 Redshifts

Finally, we show the MC predictions of the redshifts and their errors for the DESI GALAXY and QSO objects. In Fig. 23, we can see a good agreement between the predictions with the ground truth and an average redshift error (Δz) of the two classes of 2.8 per cent for galaxies and 4.8 per cent for QSOs. These errors are larger than the ones obtained for former data sets for two main reasons. The first is the unbalanced redshift distribution, especially in the high-redshift part (i.e. $z > 0.4$ for galaxies and $z > 2.5$ for QSO), where there are fewer systems, especially for the galaxies. The second is the overall smaller training samples available for these early-release data from DESI (about 1/10 of the former data sets), resulting in typically larger errors on the individual spectra. Once we can include more DESI training samples, and use customized subnetworks for the special subclasses, we expect the accuracy will rise to the level found for SDSS and 4MOST.

5 DISCUSSION

In this section, we will discuss the potential strategies for improvements in performance and further developments.

As far as classification is concerned, a key problem is how to improve the 'absolute' accuracy of the classification method. So far, we have benchmarked GaSNet-II with respect to the labels assigned from the different data sets (relative performances). For the SDSS and DESI data sets, the labels are deduced from the PCA fitting, and this can bring some systematics. In fact, when using a classification based on real spectra as labels for the training of the DL tools, the upper limit of the 'absolute' accuracy of the trained networks is decided by the accuracy of the training set, which in turn is set by the accuracy of the 'traditional' pipeline used for labelling it. A viable alternative is to incorporate human-labelled data, like, for example, SDSS-DR12 superset (Pâris et al. 2017, 2018). However, this approach is not bias-free either, introducing a different form of bias: human judgment. Another physically motivated alternative is to utilize mock data, based on theoretical templates, for example, similar to those used for the 4MOST sample in Section 4.3.1. In Fig. 24, we describe a general procedure for training on simulated data. Here, the function F represents:

$$F\ (\text{flux}) = \begin{cases} (P_i, 0), & i \in \text{galatic} \\ (P_i, z_i), & i \in \text{extragalactic}. \end{cases} \quad (13)$$

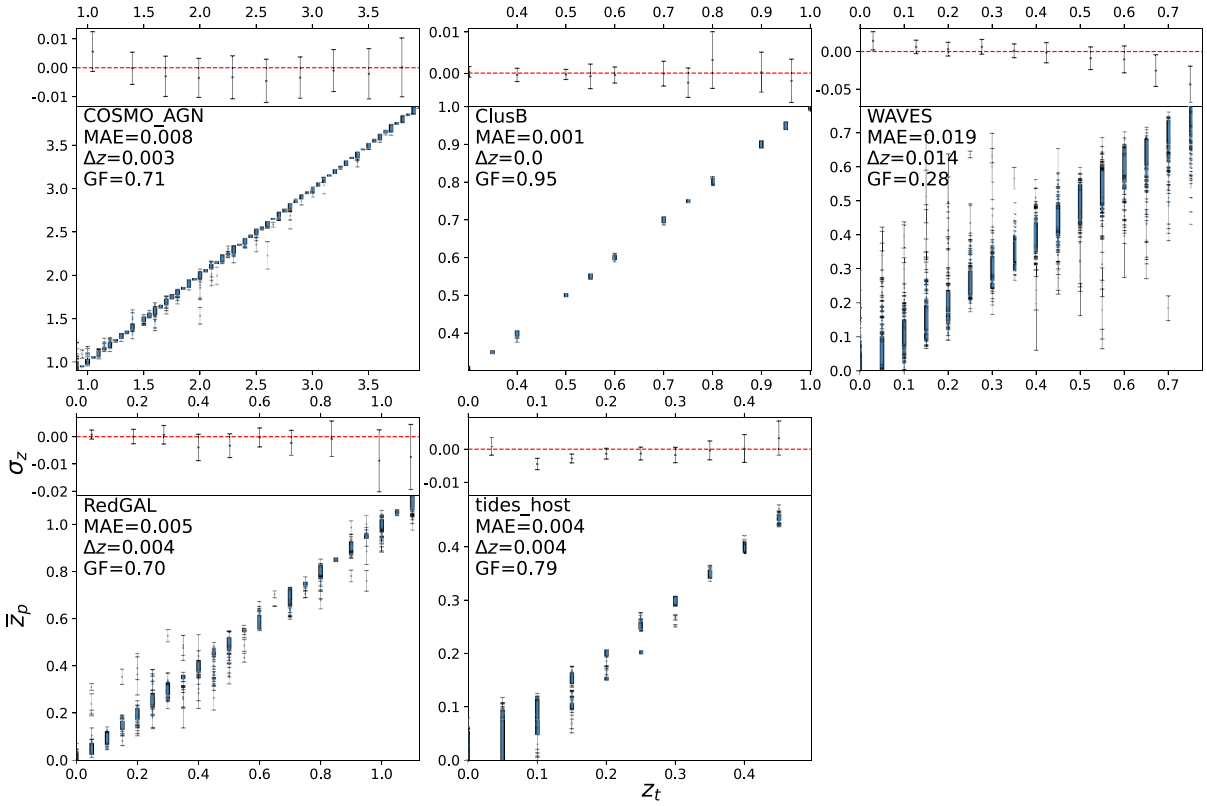

Figure 21. Redshift predictions for the five extragalactic 4MOST mock subclasses. It is worth noting that the simulated spectra are produced on a coarse grid of redshifts, hence the quantization. Legends are identical to Fig. 15.

Figure 22. The DESI classification on the test set. Legends are identical to Fig. 12. As before, the matrix should be read along columns, that is the direction along which the 100 per cent of the true labels are distributed by the classifier.

The networks shown in the figure serve as a powerful fitting tool that minimizes the need for manual adjustments. The mock data, produced under specific physical conditions (i, z_i), are used as training data for the networks. Subsequently, well-trained networks are set up by optimizing the prediction accuracy of the parameters (i, z_i). If the training sample is complete and accurate, these well-trained networks can be considered, by construction, as the optimal tools maximizing the 'absolute' accuracy of the predicting parameters (i, z_i) when applied to real observational data. In practice, this is possibly true only if: (1) the theoretical models are correct, and (2) one introduces into the process all the observational conditions to maximize the fidelity

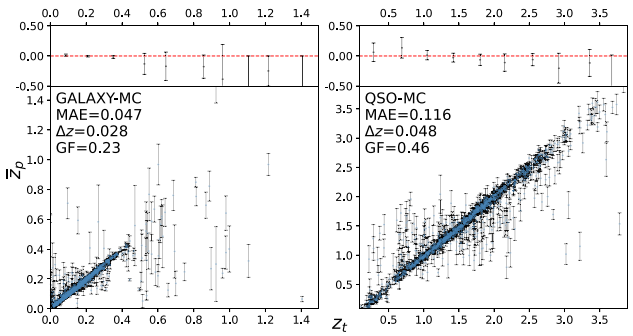

Figure 23. Redshift predictions of two DESI classes (GALAXY and QSO). Legends are identical to Fig. 15.

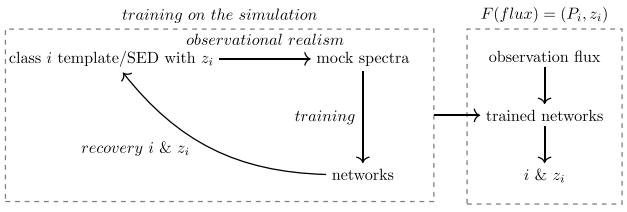

Figure 24. The general process of networks trained by simulation involves training on mock samples, finding the mapping F, and predicting real data. The training data are generated with specific parameters (i, z_i) and observational realism. The networks are trained to recover the labels i, z_i, and ultimately, the well-trained networks are used to fit the parameters i, z_i based on observational data input. It is a first-principles-based method rather than an empirical-based one.

between mock train/test sets and observations, including Poissonian noise, realistic distributions of SNR, seeing, intrinsic broadening of the features (e.g. galaxy kinematics), artefacts, etc. (see e.g. Fig. 24). The former condition is generally satisfied for most of the objects one expects to classify in galactic and extragalactic surveys as there are rather robust theoretical stellar (e.g. Coelho 2014) and galaxy/QSO templates (e.g. Kewley et al. 2001). However, there might still be remaining systematics due to specific model shortcomings or even 'unknown' phenomena that are not fully accounted for by standard theories or empirical models. In principle, these latter systems would possibly appear as 'anomalies' in theoretical-based classifications that can be studied separately either to improve models or explore new phenomena. With regard to 'observational realism', the inclusion of more observational conditions is something that is currently under development (in the case of imaging data, see e.g. Yin et al. 2022). Despite these difficulties, which we aim to address in future analyses, the main advantage of using mock data sets is the freedom to choose the hyperparameters that one is expected to predict with spectra, and then optimize the training sample accordingly (a kind of active learning loop), for example, using theoretical-based simulation spectra covering a wide and physical range of these hyperparameters. Another advantage of using the mock spectra is that they do not suffer the poor sampling problem, which plagues empirical data sets (e.g. rare events, like strong gravitational lenses, or high-redshift galaxy samples, etc.). As a result, they eliminate the biases introduced by incomplete or poor sampling.

Regardless of the philosophy behind the training sets, there might be further strategies that can help improve the classification. One is the hierarchy. Classifications can be done in one step (as we have proposed in Sections 4.2.1, 4.3.1, and 4.4.1) or multiple steps. Spectra can be roughly classified in the first step, followed by a more sophisticated subclassification in subsequent steps (see e.g. Sánchez-Sáez et al. 2021). This decision-tree-like classification can allow us to have a more fine-grained and detailed classification process. The architecture of multiple identical subnetworks, similar to what we currently use, can be easily rearranged into a decision-tree-like hierarchical structure to realize a multi-ML model combination 'tree' structure, with more branches and deeper layers.

Moving to the redshift estimates, we foresee that relevant improvements can be obtained using 'self-attention' (Vaswani et al. 2017), which is becoming popular as the state-of-the-art model in DL applications. For instance, Fig. 20 is an example where the classifier based on the ResNet struggles to effectively recognize the slight difference in the spectrum when there is a mix of features like the spectrum continuum and emission lines. 'self-attention' has shown to be superior in recognizing the global features and 'long-range correlation' compared to CNNs (Han et al. 2020) with the net effect that both classification and redshift estimates can highly be improved (see also Section 4.2.3). We plan to implement these alternative approaches in future work by replacing the convolutions with 'self-attention' in the small blocks of our network.

Finally, alternative methods of estimating the redshift error exist. In standard networks, keeping the inputs the same leads to the same outputs, which is stable but does not allow us to generalize the error estimates. Apart from introducing multiple subnetworks to estimate the errors, as we have already experimented with in this paper, there are other approaches to introduce uncertainty, such as MC dropout techniques (Podsztavek, Škoda & Tvrdík 2022) or Bayesian neural networks (Perreault Levasseur, Hezaveh & Wechsler 2017; Zhou et al. 2022; Gentile et al. 2023). We stress though that we expect that these methods are unlikely to yield significant differences with respect to our approach as these methods obtain the error by repeating predictions. We aim to test these different techniques in future analyses.

6 CONCLUSION

We have developed new tools for spectroscopy classification and redshift prediction using DL techniques and constructed a pipeline that we have tested on SDSS, 4MOST, and DESI data sets. The performance of our pipeline on these three different data sets can be summarized as follows: on SDSS, the classifier achieves an average accuracy of 92.4 per cent for a 13-subclass classification task (with most types exceeding 90 per cent), and redshift prediction accuracy around 0.23 per cent for galaxy and 2.1 per cent for QSO subclasses. On 4MOST, the classifier achieves an average accuracy of 93.4 per cent for a 10-subclass classification task and redshift prediction accuracy of around 0.55 per cent for galaxy and 0.3 per cent for AGN. On DESI, the classifier achieves an average accuracy of 96 per cent for a 3-class classification task and redshift prediction accuracy of around 2.8 per cent for galaxy and 4.8 per cent for AGN. The accuracy of classifiers is strikingly consistent. However, the aspect of redshift prediction is clearly dependent on various factors such as the types of subclasses/classes, the average spectral element SNR, and the sample size of the training data. For example, the poor SNR of subclass WAVES results in the highest error on the 4MOST data set, while the relatively sparse training data for DESI contributes to a larger redshift error compared to SDSS and 4MOST.

GaSNet-II's efficiency and accuracy make this tool suitable for real-time analyses of nightly observations. The predictions for 39 000 spectra can be completed in less than one minute. Among the data products, GaSNet-II can provide realistic redshift errors from a built-in subnetwork architecture simulating an MC test. As seen in the discussion of the SDSS-DR16 results. The redshift error of each data point can be also used to assess the robustness of the predicted redshifts.

In summary, DL methods offer significant advantages for Stage-IV spectroscopic infrastructures like DESI, 4MOST, and MOONS in various aspects, such as efficiency, 'data-driven', better performance in low SNR, better consistency and systematics, and so on. Although the current redshift accuracy leaves room for improvement, DL, as a new tool, holds huge potential for further development. Many aspects of improvement can be done with the future 4MOST simulations. Further data sets such as theoretical spectra and improvements such as a 'self-attention' structure will be applied to GaSNet-II in the future to improve the 'absolute' accuracy of classification and redshift estimates, respectively.

ACKNOWLEDGEMENTS

We thank D. De Martino and P. Szkody for their useful comments and suggestions. NRN and FZ acknowledge that part of this work was supported by the National Science Foundation of China, the Research Fund for Excellent International Scholars (grant no. 12150710511), and the research grant from China Manned Space Project no. CMS-CSST-2021-A01. FZ acknowledges the support of the China Scholarship Council (grant n. 202306380249). CH's work is funded by the Volkswagen Foundation. CH acknowledges additional support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1-390900948 (the Heidelberg STRUCTURES Excellence Cluster). ML acknowledges financial support from INAF-Minigrant "4MOST-StePS: a Stellar Population Survey using 4MOST@VISTA" (2022). CR acknowledges support from Fondecyt Regular grant 1230345

and ANID BASAL project FB210003. LC acknowledges support by grants PIB2021-127718NB-100 and PID2022-139567NB-I00 from the Spanish Ministry of Science and Innovation/State Agency of Research MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe". RL acknowledges the support of the National Nature Science Foundation of China (No 12203050). This work was funded by ANID – Millennium Science Initiative Program – ICN12_009 (FEB), CATA-BASAL – FB210003 (FEB, BLR, and CS), and FONDECYT Regular – 1200495 (FEB and BLR). BFR acknowledges the Polish LSST/Rubin grant from the Ministry of Science and Higher Education (MNiSW) DIR/WK/2018/12. Y-LK has received funding from the Science and Technology Facilities Council (grant no. ST/V000713/1). RJA was supported by FONDECYT grant no. 1231718 and by the ANID BASAL project FB210003. LJMD acknowledges funding by the Australian Research Council (ARC) Future Fellowship scheme (FT200100055). GG acknowledges support by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project-IDs: eBer-22-59652 (GU 2240/1-1). This project has also received additional funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 949173). ET acknowledges the ETAg grant PRG1006 and the CoE project TK202 funded by the HTM. The project that gave rise to these results received the support of a fellowship from the ``la Caixa" Foundation (ID 100010434). The fellowship code is LCF/BQ/PR24/12050015.

DATA AVAILABILITY

The code and data (SDSS) are available in the GitHub link: https://github.com/Fucheng-Zhong/GaSNet-II.

REFERENCES

Ahumada R. et al., 2020, ApJS, 249, 3
Alexander D. M. et al., 2023, AJ, 165, 124
Almeida A. et al., 2023, ApJS, 267, 44
Alzubaidi L. et al., 2021, J. Big Data, 8, 1
Ball N. M., Loveday J., Fukugita M., Nakamura O., Okamura S., Brinkmann J., Brunner R. J., 2004, MNRAS, 348, 1038
Bellstedt S. et al., 2020, MNRAS, 498, 5581
Bensby T. et al., 2019, The Messenger, 175, 35
Bernardi M., Hyde J. B., Sheth R. K., Miller C. J., Nichol R. C., 2007, AJ, 133, 1741
Bialek S., Fabbro S., Venn K. A., Kumar N., O'Briain T., Yi K. M., 2020, MNRAS, 498, 3817
Bolton A. S. et al., 2012, AJ, 144, 144
Boucaud A. et al., 2020, MNRAS, 491, 2481
Bundy K. et al., 2015, ApJ, 798, 7
Busca N., Balland C., 2018, preprint (arXiv:1808.09955)
Chiappini C. et al., 2019, The Messenger, 175, 30
Christlieb N. et al., 2019, The Messenger, 175, 26
Cid Fernandes R., Mateus A., Sodré L., Stasińska G., Gomes J. M., 2005, MNRAS, 358, 363
Cioni M.. R. L. et al., 2019, The Messenger, 175, 54
Cirasuolo M. et al., 2020, The Messenger, 180, 10
Coelho P. R. T., 2014, MNRAS, 440, 1027
Comparat J. et al., 2020, A&A, 636, A97
DESI Collaboration, 2022, AJ, 164, 207
DESI Collaboration, F. Edward 2023, AJ 166 259 preprint(arXiv:2306.06308)
D'Isanto A., Polsterer K. L., 2018, A&A, 609, A111
Dawson K. S. et al., 2016, AJ, 151, 44
Domínguez Sánchez H., Margalef B., Bernardi M., Huertas-Company M., 2022, MNRAS, 509, 4024
Driver S. P. et al., 2019, The Messenger, 175, 46
Fabbro S., Venn K. A., O'Briain T., Bialek S., Kielty C. L., Jahandar F., Monty S., 2018, MNRAS, 475, 2978
Finoguenov A. et al., 2019, The Messenger, 175, 39
Fiore F. et al., 2017, A&A, 601, A143
Fitzpatrick E. L., Massa D., 2007, ApJ, 663, 320
Ganaie M. A., Hu M., Malik A. K., Tanveer M., Suganthan P. N., 2022, Engineering Applications of Artificial Intelligence 115 105151 preprint(arXiv:2104.02395)
Gentile F., Tortora C., Covone G., Koopmans L. V. E., Li R., Leuzzi L., Napolitano N. R., 2023, MNRAS, 522, 5442
Graff P., Feroz F., Hobson M. P., Lasenby A., 2014, MNRAS, 441, 1741
Guiglion G. et al., 2024, A&A, 682, A9
Guo Z., Martini P., 2019, ApJ, 879, 72
Han K. et al., 2023, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 87–110 preprint(arXiv:2012.12556)
He K., Zhang X., Ren S., Sun J., 2016, Proceedings of the IEEE Conference on CVPR 770–778 preprint(arXiv:1512.03385)
Helmi A. et al., 2019, The Messenger, 175, 23
Hoyle B., 2016, Astron. Comput., 16, 34
Huang X. et al., 2020, ApJ, 894, 78
Hutchinson T. A. et al., 2016, AJ, 152, 205
Jin S. et al., 2024, MNRAS, 530, 2688
Jönsson H. et al., 2020, AJ, 160, 120
Kewley L. J., Dopita M. A., Sutherland R. S., Heisler C. A., Trevena J., 2001, ApJ, 556, 121
Kim E. J., Brunner R. J., 2017, MNRAS, 464, 4463
Krizhevsky A., Sutskever I., Hinton G. E., 2012, Commun. ACM, 60, 84
Lakshminarayanan B., Pritzel A., Blundell C., 2017, Advances in neural information processing systems 30 preprint(arXiv:1612.01474)
Lan T.-W. et al., 2023, ApJ, 943, 68
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Lehnert M. D., Heckman T. M., 1996, ApJ, 472, 546
Leung H. W., Bovy J., 2019, MNRAS, 483, 3255
Li R., Shu Y., Su J., Feng H., Zhang G., Wang J., Liu H., 2019, MNRAS, 482, 313
Li R. et al., 2021, ApJ, 923, 16
Li R. et al., 2022a, A&A, 666, A85
Li R., Napolitano N. R., Roy N., Tortora C., La Barbera F., Sonnenfeld A., Qiu C., Liu S., 2022b, ApJ, 929, 152
Liu H.-Y., Liu W.-J., Dong X.-B., Zhou H., Wang T., Lu H., Yuan W., 2019, ApJS, 243, 21
Lyke B. W. et al., 2020, ApJS, 250, 8
Makhija S., Saha S., Basak S., Das M., 2019, Astron. Comput., 29, 100313
Mateus A., Sodré L., Cid Fernandes R., Stasińska G., Schoenell W., Gomes J. M., 2006, MNRAS, 370, 721
Merloni A. et al., 2019, The Messenger, 175, 42
Nepal S. et al., 2023, A&A, 671, A61
Pâris I. et al., 2017, A&A, 597, A79
Pâris I. et al., 2018, A&A, 613, A51
Parks D., Prochaska J. X., Dong S., Cai Z., 2018, MNRAS, 476, 1151
Pasquet J., Bertin E., Treyer M., Arnouts S., Fouchez D., 2019, A&A, 621, A26
Perreault Levasseur L., Hezaveh Y. D., Wechsler R. H., 2017, ApJ, 850, L7
Petrillo C. E. et al., 2019, MNRAS, 482, 807
Podsztavek O., Škoda P., Tvrdík P., 2022, Astron. Comput., 40, 100615
Richard J. et al., 2019, The Messenger, 175, 50
Robotham A. S. G., Bellstedt S., Lagos C. d. P., Thorne J. E., Davies L. J., Driver S. P., Bravo M., 2020, MNRAS, 495, 905
Sánchez-Sáez P. et al., 2021, AJ, 161, 141
Secrest N. J., von Hausegger S., Rameez M., Mohayaee R., Sarkar S., Colin J., 2021, ApJ, 908, L51
SubbaRao M., Frieman J., Bernardi M., Loveday J., Nichol B., Castander F., Meiksin A., 2002, in Starck J.-L., Murtagh F. D.eds, Proc. SPIE Conf. Ser. Vol. 4847, Astronomical Data Analysis II. SPIE, Bellingham, p. 452,
Swann E. et al., 2019, The Messenger, 175, 58

Talbot M. S., Brownstein J. R., Dawson K. S., Kneib J.-P., Bautista J., 2021, MNRAS, 502, 4617
Thorne J. E. et al., 2021, MNRAS, 505, 540
Tonry J., Davis M., 1979, AJ, 84, 1511
Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I., 2017, Advances in neural information processing systems 30 preprint(arXiv:1706.03762)
Yan R. et al., 2019, ApJ, 883, 175
Yin J. E., Eisenstein D. J., Finkbeiner D. P., Protopapas P., 2022, PASP, 134, 044502
Zhan H., 2011, Sci. Sin. Phys. Mech. Astron., 41, 1441
Zhong F., Li R., Napolitano N. R., 2022, Res. Astron. Astrophys., 22, 065014
Zhou X., Gong Y., Meng X.-M., Chen X., Chen Z., Du W., Fu L., Luo Z., 2022, Res. Astron. Astrophys., 22, 115017
de Diego J. A. et al., 2020, A&A, 638, A134
de Jong R. S. et al., 2019, The Messenger, 175, 3

APPENDIX A: CROSS-CONTAMINATION ON DATA SET

We have anticipated in Section 2.1 that the empirical classification of the SDSS-DR16 cannot guarantee full accuracy, and we cannot exclude cross-contamination among the different classes. This might have an impact both on the classification and the accuracy of the redshifts. As discussed in Section 5, a possible workaround is to train on a purer sample of mock spectra based on well-established theoretical or observational templates (Bellstedt et al. 2020; Robotham et al. 2020; Thorne et al. 2021). An example of how this might lead to higher performances has been offered by the 4MOST sample, where for some classes we have reached 100 per cent accuracy (e.g. COSMOS_AGN, ESN, and GAL_HR) for a combination of clean templates and rich training sample, although the 4MOST training sample is not exactly built over physically motivated templates, but, rather, specific survey targets, that might have very specific properties, including high SNR, that make the spectra easier to classify (e.g. tides_host).

Here, we intend to check, more quantitatively, the possible impact of the misclassified spectra in a given class. We use the SDSS-DR16 data set as a reference for this test and add, to each class, 5 per cent or 10 per cent contamination from the relatively similar classes seen from the confusion matrix. For the sake of brevity and clarity, we use only three extragalactic classes: AGN, STARBURST, and STARFORMING. The result is shown in the confusion matrix in Fig. A1. All of those three classes belong to the subclass of GALAXY and show some degree of mixing with each other in the SDSS sample.

As additional testing, the results of cross-contamination on the 4MOST data set (RedGAL, clusB, and COSMO_AGN) are also shown in Fig. A2. In Section 4.3.1, the confusion matrix shows some strong degeneracy between the class RedGAL and clusB, so we test the cross-contamination on those two classes and COSMO_AGN. The additional class COSMO_AGN was used to reflect the upper limit of classification (the cross-contamination rate). We find an average decrease in accuracy by 5.7 per cent for 5 per cent contamination and 9 per cent for 10 per cent contamination.

We end this section by showing the redshift predictions for the samples with contamination, discussed at the end of the Section 4.3.1. Fig. A3 shows the redshift prediction results for three subclasses (AGN, STARBURST, and STARFORMING) in the three different cross-contamination levels (0 per cent – 5 per cent – 10 per cent). The figure shows that the small contamination among subclasses does not significantly decrease the accuracy of redshift prediction, but it still causes performance degradation, such as the average decrease in GF by 2 per cent for 5 per cent contamination and 3.3 per cent for 10 per cent contamination.

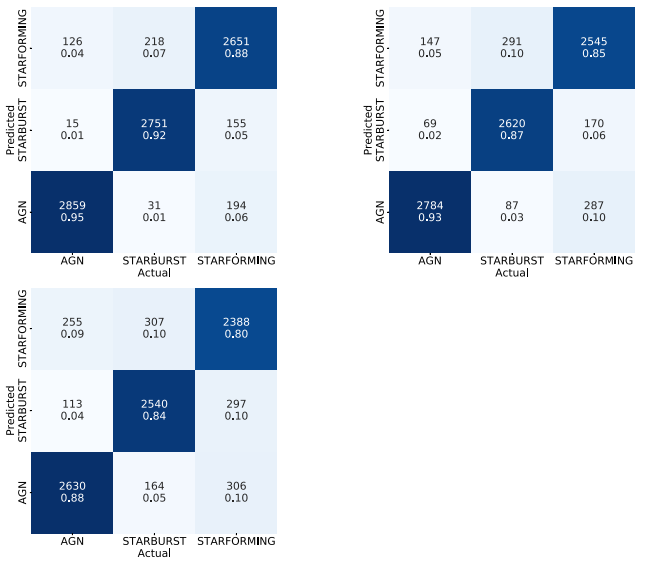

Figure A1. Confusion matrix of a three-class classification (AGN, STARBURST, and STARFORMING), showing how this changes with increasing random contamination (0 per cent – 5 per cent – 10 per cent). The contamination fraction refers to randomly selected and shuffled labels in the data set. As expected, as contamination increases, the accuracy decreases. Roughly speaking, with respect to the original sample (0 per cent artificial contamination), we observe an average decrease in accuracy by 3.3 per cent for 5 per cent contaminants and 7.7 per cent for 10 per cent contaminants.

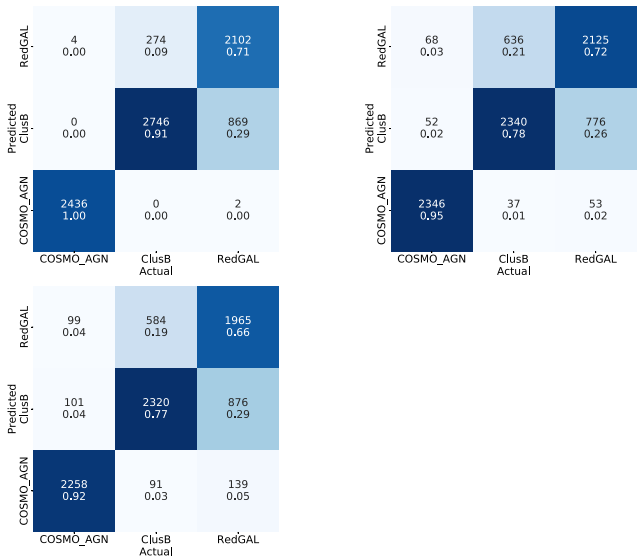

Figure A2. Confusion matrix of a three-class classification (COMOS_AGN, ClusB, and RedGAL), showing how this changes with increasing random contamination (0 per cent – 5 per cent – 10 per cent). The contamination fraction refers to randomly selected and shuffled labels in the data set. As contamination increases, the accuracy of both the COMOS_AGN and ClusB decreases. Roughly speaking, with respect to the original sample (0 per cent artificial contamination), we observe an average decrease in accuracy by 5.7 per cent for 5 per cent contamination and 9 per cent for 10 per cent contamination.

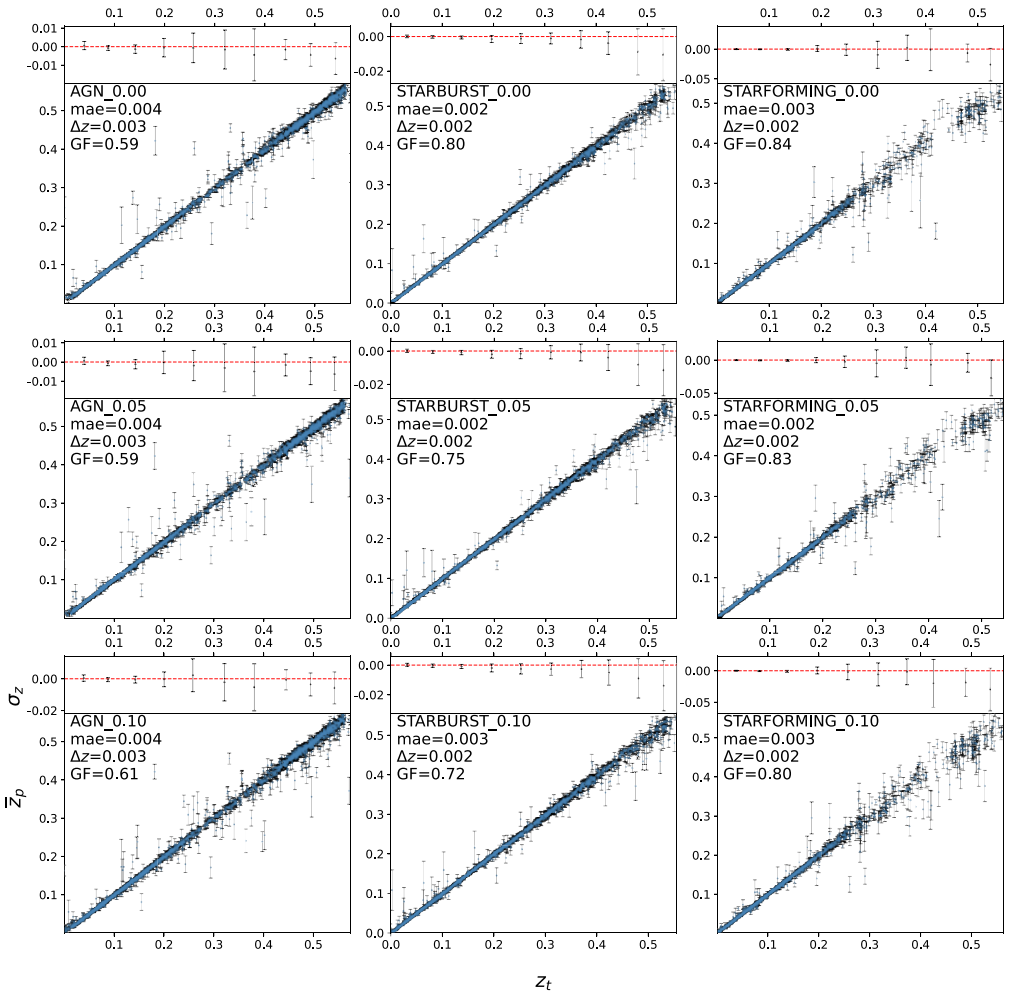

Figure A3. Redshift prediction results of three subclass (AGN, STARBURST, and STARFORMINGL) in the three different cross-contamination levels (0 per cent – 5 per cent – 10 per cent). The first row represents 0 per cent contamination. The second row represents 5 per cent contamination. The third row represents 10 per cent contamination. The figure indicates that the small contamination on subclasses does not significantly decrease the accuracy of redshift prediction, but we still observe an average decrease of GF by 2 per cent for 5 per cent contamination and 3.3 per cent for 10 per cent contamination.

APPENDIX B: THE COARSE CLASSIFIER OF SDSS-DR16 AND 4MOST

In this appendix, we briefly describe a more homogeneous comparative check of the performances of the GaSNet-II on the three data sets discussed in the paper, by emulating the situation where we have the same data size and number of classes. We use the DESI sample as a reference, as it contains the smaller data set (21 000 entries) and coarser classification (GALAXY, QSO/AGN, and STAR). To do that, we have regrouped the spectra belonging to these three broader classes for SDSS (STAR: raw 1–7; GALAXY: raw 8–11; and QSO: raw 12–13, in Table 1) and 4MOST (STAR: raw 1–5; AGN: raw 6; and GALAXY: raw 7–10, in Table 2) respectively. To be uniform with the DESI case, we have also randomly extracted 7000 spectra from these re-grouped classes, to train and test the GaSNet-II, using the same set-up of DESI training/testing. Fig. B1 shows the results of this 'coarse' classification of SDSS and 4MOST data sets, to be compared with the same for DESI in Fig. 22. The figure indicates that GaSNet-II can achieve an average accuracy of 99 per cent for classification. The STAR class nearly achieved 100 per cent accuracy. The class with the lowest accuracy is QSO, but it still achieved an impressive 98 per cent accuracy. Compared to the DESI classification, it exhibits a higher accuracy for the same coarse classes and the same amount of training data, with an improvement of about 3 per cent, which can be attributed to the qualities of the SDSS spectrum. For instance, the

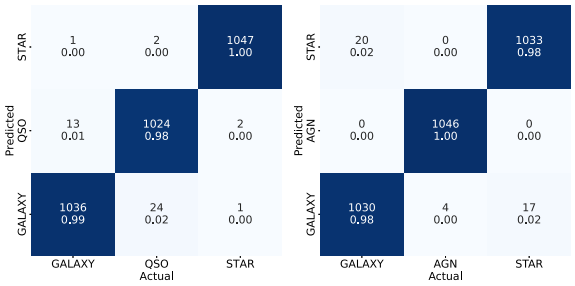

Figure B1. Results of the 'coarse' classification of the SDSS (left) and 4MOST (right) data set. The spectra are categorized into three classes: GALAXY, QSO (AGN), and STAR. GaSNet-II achieved an average accuracy of 99 per cent. The STAR class nearly achieved 100 per cent accuracy. The class with the lowest accuracy is QSO, but it still achieved an impressive 98 per cent accuracy.

mean SNR of stars and galaxies in SDSS spectra is higher than that of DESI. This can be seen by comparing Tables 1 and 3.

APPENDIX C: RELATIONSHIP BETWEEN AVERAGE CLASSIFICATION ACCURACY AND SNR

Here, we want to test the dependence of the classification accuracy

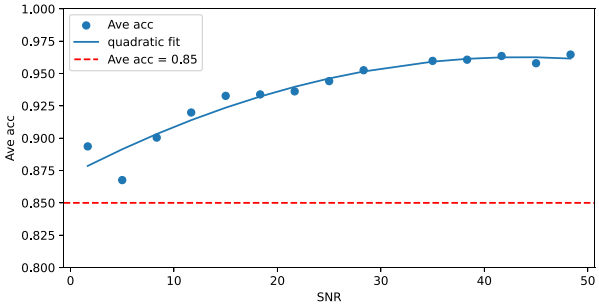

Figure C1. The average classification accuracy of the SDSS 13-subclasses classification with respect to the SNR. We only consider the SNR range of 0–50.

on the SNR of the spectra (see also Section 4.2.3 for the redshift estimates). In Fig. C1, we show the average classification accuracy over the SDSS 13-subclasses with respect to the SNR. We consider 14 bins in the SNR range of 0–50. The figure shows that as the SNR increases, the accuracy systematically increases and finally reaches an upper limit of an average classification accuracy of $\sim$ 96 per cent.

APPENDIX D: VISUALIZATION, THE GRADIENTS OF OUTPUT

As discussed in Nepal et al. 2023, target (i.e. output label) gradients as a function of input neuron (or wavelength), in the form of partial derivatives of the output with respect to λ can give information about the sensitivity of output labels to each of the input fluxes. This allows us to visualize whether the CNN is learning from the spectral features. In Fig. D1, we have selected six SDSS extragalactic random spectra including objects from different classes. We have paid attention to avoiding too low SNR to avoid the gradient being dominated by noise rather than the impact of the spectral features. As it can be seen, the gradients of both the classification probability, $|\frac{\mathrm{d}P}{\mathrm{d}\lambda}|$, and the redshift predictions, $|\frac{\mathrm{d}z}{\mathrm{d}\lambda}|$, show strong increases around the most prominent features (e.g. emission lines in star-forming and starburst galaxies) and possibly some absorption lines from normal galaxies. Interestingly, they seem to be less sensitive to the very broad lines from quasars, meaning that these are too smoothly varying, maybe looking more like a continuum feature. Also interesting is the fact that the gradients show a burst around the 'redshifted' 4000 Å break for the GALAXY_nan spectrum (at $\sim$7000 Å), implying that this is a feature that can be seen by the CNN.

APPENDIX E: CLASSIFICATION ACCURACY, REDSHIFT UNCERTAINTY, AND VELOCITY DISPERSION

In this appendix, we test the impact of the VDISP on the spectra classification and redshift estimates. The line broadening caused by the VDISP might enlarge the width of emission or absorption lines,

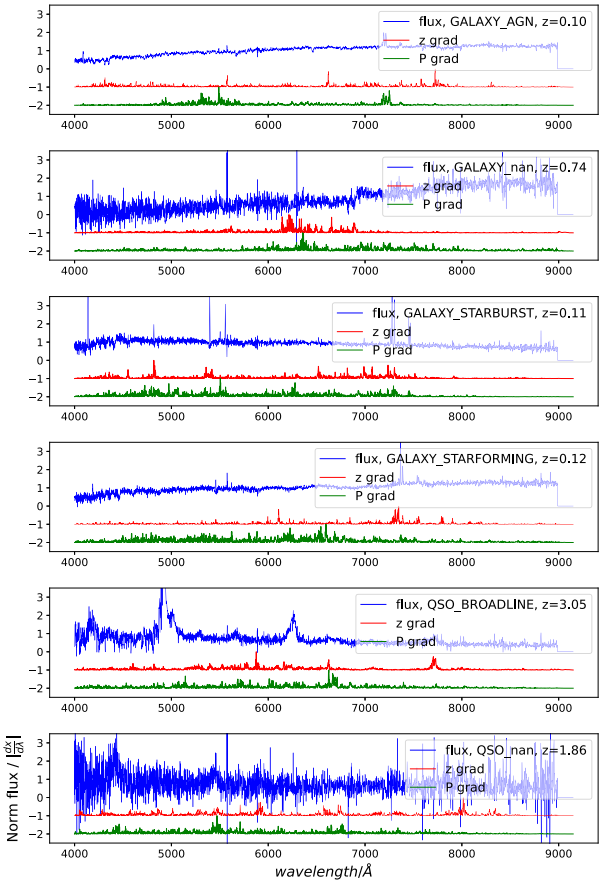

Figure D1. The normalized flux and gradients of six SDSS extragalactic spectra. Line z_grad represents the absolute redshift gradients, $|\frac{\mathrm{d}z}{\mathrm{d}\lambda}|$, which is shifted by -1; and line P_grad represents the absolute probability gradients, $|\frac{\mathrm{d}P}{\mathrm{d}\lambda}|$, which is shifted by -2.

affecting the accuracy of redshift prediction. In Fig. E1 (left panel), we demonstrate that the predicted σ_z of the SDSS data set is slightly correlated with the VDISP of galaxies, as the larger the VDISP of the galaxy, the larger the predicted uncertainty. However, in the same figure, we also show the σ_z separated in the different subclasses and we see that, for each subclass, the σ_z is almost independent of the VDISP. This is mirrored by the classification accuracy (bottom panel) where we see that, except for 'GALAXY_STARFORMING' which has a sparser sampling, the accuracy also stays almost constant with the VDISP. Hence, we conclude that the accuracy of classifications and redshift estimates are mainly driven by the class type (meaning spectral features) and SNR (see Section 4.2.3 and Appendix C), rather than the VDISP.

[1]*School of Physics and Astronomy, Sun Yat-sen University, Zhuhai Campus, 2 Daxue Road, Xiangzhou District, 519082, Zhuhai, P. R. China*
[2]*Department of Physics E. Pancini, University Federico II, Via Cinthia 6, I-80126, Naples, Italy*
[3]*Institute of Theoretical Physics, University of Heidelberg, Philosophenweg 12, D-69120, Heidelberg, Germany*
[4]*Institude for Astrophysics, School of Physics, Zhengzhou University, Zhengzhou, 450001, China*
[5]*Instituto de Astrofísica and Centro de Astroingeniería, Facultad de Física, Pontificia Universidad Católica de Chile, Campus San Joaquín, Av. Vicuña Mackenna 4860, Macul Santiago 7820436, Chile*
[6]*Millennium Institute of Astrophysics, Nuncio Monseñor Sótero Sanz 100, Of 104, Providencia, 7500011, Santiago, Chile*

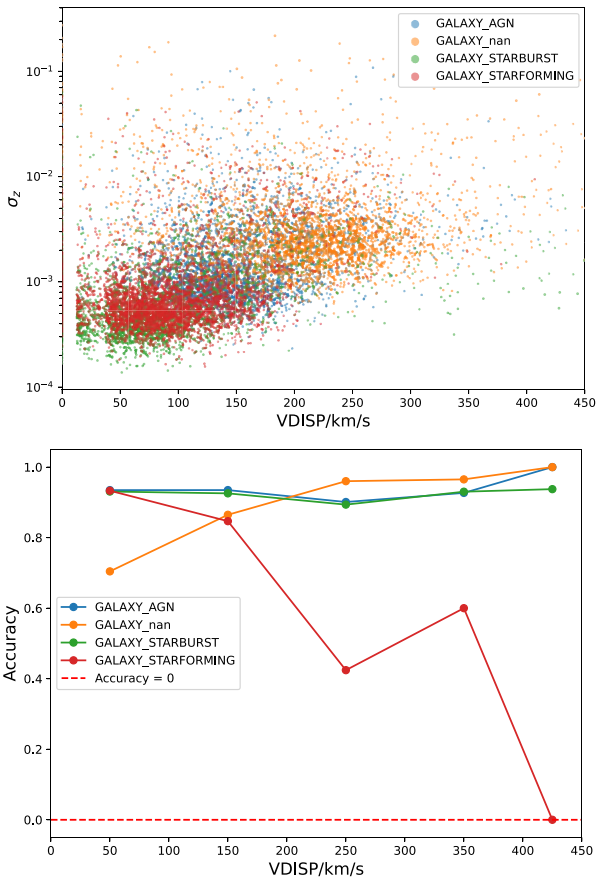

Figure E1. The predicted σ_z values by the MC and the VDISP of four SDSS subclasses of galaxies are plotted. The x-axis represents the VDISP, limited to values up to 450 km s^{-1}.

[7]*Space Science Institute, 4750 Walnut Street, Suite 205, Boulder, CO 80301, USA*
[8]*Centre National de la Recherche Scientifique (CNRS), Centre of Research in Astrophysics of Lyon (CRAL)Université de Lyon, 9 av Charles André, F-69230 Saint Genis Laval,, France*
[9]*Max-Planck-Institut für Extraterrestrische Physik (MPE), Giessenbach-strasse 1, D-85748 Garching bei München, Germany*
[10]*Department of Physics, Lancaster University, Lancaster LA1 4YB, UK*
[11]*Centre de Recherche Astrophysique de Lyon, 9 Avenue Charles André, F-69230 Saint-Genis-Laval, France*
[12]*INAF – Osservatorio Astronomico di Brera, via Brera 28, I-20121 Milano, Italy*
[13]*Astronomy Centre, University of Sussex, Falmer, Brighton BN1 9QH, UK*
[14]*Institute of Astronomy, Faculty of Physics, Astronomy and Informatics, Nicolaus Copernicus, University, Grudziadzka 5, PL-87-100 Toruń, Poland*
[15]*Centre de Recherche Astrophysique de Lyon UMR5574 , Univ Lyon, Ens de Lyon, Univ Lyon1, CNRS, F-69007 Lyon, France*
[16]*Max-Planck-Institut für Extraterrestrische Physik (MPE), Gießenbach-straße 1, D-85748 Garching bei München, Germany*
[17]*INAF – Osservatorio Astronomico di Capodimonte, Salita Moiariello 16, I-80131 Napoli, Italy*
[18]*Instituto de Estudios Astrofísicos, Facultad de Ingeniería y Ciencias, Universidad Diego Portales, Av. Ejército 441, Santiago, Chile*
[19]*INAF – IASF Milano, via A. Corti 12, I-20133 Milano, Italy*
[20]*Centro de Astrobiología (CAB), CSIC-INTA, Ctra. de Ajalvir km 4, Torrejón de Ardoz, E-28850 Madrid, Spain*
[21]*Sydney Institute for Astronomy, School of Physics, University of Sydney, NSW 2006, Australia*
[22]*ICRAR, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia*
[23]*OmegaLambdaTec GmbH, Parkring 6, D-85748 Garching, Germany*
[24]*Zentrum für Astronomie der Universität Heidelberg, Landessternwarte, Königstuhl 12, D-69117 Heidelberg, Germany*
[25]*Max Planck Institute for Astronomy, Königstuhl 17, D-69117 Heidelberg, Germany*
[26]*Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, D-14482 Potsdam, Germany*
[27]*Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, Porto, 4150-762, Portugal*
[28]*DTx – Digital Transformation CoLab, Building 1, Azurém Campus, University of Minho, P-4800-058 Guimarães, Portugal*
[29]*European Southern Observatory, Science Operations, Alonso de Cordova 3107, Vitacura, 19001 Santiago, Chile*
[30]*Kavli Institute for Astronomy and Astrophysics, Peking University, Beijing 100871, China.*
[31]*Instituto de Física, Pontificia Universidad Católica de Valparaíso, Av. Universidad 330, Curauma, 2373223, Valparaíso, Chile*
[32]*Tartu Observatory, University of Tartu, Observatooriumi 1, 61602 Tõravere, Estonia*
[33]*Estonian Academy of Sciences, Kohtu 6, 10130 Tallinn, Estonia*
[34]*School of Mathematical and Physical Sciences, Macquarie University, NSW 2109, Australia*

This paper has been typeset from a TEX/LATEX file prepared by the author.