Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **502,** 1355–1365 (2021)
Advance Access publication 2021 January 8

doi:10.1093/mnras/stab030

Constraining stellar population parameters from narrow band photometric surveys using convolutional neural networks

Choong Ling Liew-Cain ,[1★] Daisuke Kawata ,[1] Patricia Sánchez-Blázquez,[2,3] Ignacio Ferreras [1,4,5] and Myrto Symeonidis[1]

[1] *Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking RH5 6NT, UK*
[2] *Departamento de Física de la Tierra y Astrofísica, Universidad Complutense de Madrid, E-28040 Madrid, Spain*
[3] *Instituto de Física de Partículas y del Cosmos IPARCOS, Facultad de Ciencias Físicas, Universidad Complutense de Madrid, E-28040 Madrid, Spain*
[4] *Instituto de Astrofísica de Canarias, Calle Vía Láctea s/n, E-38205 La Laguna, Tenerife, Spain*
[5] *Departamento de Astrofísica, Universidad de La Laguna (ULL), E-38206 La Laguna, Tenerife, Spain*

Accepted 2020 December 23. Received 2020 December 17; in original form 2020 February 19

ABSTRACT

Upcoming large-area narrow band photometric surveys, such as Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS), will enable us to observe a large number of galaxies simultaneously and efficiently. However, it will be challenging to analyse the spatially resolved stellar populations of galaxies from such big data to investigate galaxy formation and evolutionary history. We have applied a convolutional neural network (CNN) technique, which is known to be computationally inexpensive once it is trained, to retrieve the metallicity and age from J-PAS-like narrow-band images. The CNN was trained using synthetic photometry from the integral field unit spectra of the Calar Alto Legacy Integral Field Area survey and the age and metallicity obtained in a full spectral fitting on the same spectra. We demonstrate that our CNN model can consistently recover age and metallicity from each J-PAS-like spectral energy distribution. The radial gradients of the age and metallicity for galaxies are also recovered accurately, irrespective of their morphology. However, it is demonstrated that the diversity of the data set used to train the neural networks has a dramatic effect on the recovery of galactic stellar population parameters. Hence, future applications of CNNs to constrain stellar populations will rely on the availability of quality spectroscopic data from samples covering a wide range of population parameters.

Key words: methods: data analysis – techniques: photometric – surveys – galaxies: evolution – galaxies: fundamental parameters.

1 INTRODUCTION

The determination of the stellar population properties in galaxies is one of the most powerful techniques to understand the formation and evolution of galaxies. Traditionally, this has been done by comparing the absorption line spectral features with stellar population synthesis models (e.g. Worthey 1994; Bruzual & Charlot 2003; Vazdekis et al. 2010; Conroy 2013), using spectral indices (e.g. Trager et al. 2000; Sánchez-Blázquez 2016) or, more recently, using full spectral fitting techniques (Panter, Heavens & Jimenez 2003).

Over the past few years, galactic spectra have been obtained by integral field unit (IFU) surveys, including Calar Alto Legacy Integral Field Area (CALIFA; Sánchez et al. 2012), Mapping Nearby Galaxies at APO (MaNGA; Bundy et al. 2015), Sydney-Australian-Astronomical-Observatory Multi-object Integral-Field spectrograph (SAMI; Croom et al. 2012), *K*-band Multi Object Spectrograph (KMOS; Wisnioski et al. 2015). These IFU surveys can be used to produce two-dimensional distributions of age and metallicity to be studied for different galaxy types. These spatially resolved spectra have put strong constraints on galaxy formation and stellar population synthesis models (e.g. Belfiore et al. 2019).

An alternative to spectroscopic surveys comes from narrow band filter imaging. Photometric surveys are more efficient at observing fainter objects than spectroscopic instruments, and can cover a greater area on the sky in a single observation. In photometric surveys, galaxies are not pre-selected, unlike in spectroscopic surveys. Instead, all galaxies that are brighter than the limiting magnitude in the field of view will be observed. Narrow and medium band filter surveys, such as Classifying Objects by Medium-Band Observations-17 (COMBO 17; Wolf, Meisenheimer & Röser 2001), Survey for High-*z* Absorption Red and Dead Sources (SHARDS; Pérez-González et al. 2013), Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS; Benitez et al. 2014), Javalambre Photometric Local Universe Survey (J-PLUS; Cenarro et al. 2019), and Southern Photometric Local Universe Survey (S-PLUS; Mendes de Oliveira et al. 2019), effectively act as low spectral resolution IFU surveys, producing spectral energy distributions (SEDs) at many positions within the galaxy. These SEDs contain enough information to derive an average stellar age and metallicity (e.g. San Roman et al. 2018). For example, Díaz-García et al. (2015) used Advanced Large Homogeneous Area Medium Band Redshift Astronomical Survey (ALHAMBRA) data to derive redshift, metallicity, and age and

[★] E-mail: choongling.liew-cain.18@ucl.ac.uk compare these values with spectroscopic observations of the same galaxies. The Multi-Filter Fitting for stellar population diagnostics (MUFFIT; Díaz-García et al. 2015) code they developed shows good recovery of the spectroscopic values, though results are highly dependent on the choice of stellar population model. San Roman et al. (2019) analyse two elliptical galaxies observed by J-PLUS. The radial gradients for age, metallicity, and extinction that are derived are in reasonable agreement with CALIFA survey observations of the same galaxies.

A challenge emerging from narrow-band surveys is the volume of data to be analysed. For example, J-PAS aims to observe a total of 9×10^7 galaxies with multiple pixels per galaxy. Additionally, J-PAS and J-PLUS together are expected to collect a maximum of 1.5 TB of data per night (Benitez et al. 2014). Therefore, a computationally efficient method for deriving stellar population parameters from the data is required, and will become invaluable in the future with larger surveys. In this paper, we present neural networks as a tool that shows promise in overcoming this challenge.

Neural networks are algorithms that allow non-linear mapping between input and target parameters, and are efficient methods of analysing large data sets. Supervised machine learning uses an input data set, such as photometric SEDs, and the set of ‘true’ values of the target parameter, e.g. age or metallicity, to learn how to make accurate predictions. Selecting an appropriate training set is a vital step in neural network methods. Galaxies have a diverse formation history and therefore the training set needs to cover this wide variety of galaxy evolution. Otherwise, the neural network will not be capable of accounting for the diversity present in galactic surveys.

Machine learning is applied widely in astrophysical research (e.g. Folkes, Lahav & Maddox 1996; Baron 2019) and has been used to derive the metallicity of galaxies from broad-band photometric surveys previously. Acquaviva (2016) and Wu & Boada (2019) applied random forest algorithms and neural networks, respectively, to calculate the metallicity of galaxies from multiwavelength Sloan Digital Sky Survey (SDSS) photometric observations, with SDSS spectral age and metallicities used as training data. Lovell et al. (2019) used the results of cosmological simulations of galaxies to synthesize SDSS-like spectra. The authors included simulated effects of extinction and noise when creating these SEDs. Convolutional neural networks (CNNs) were trained on these SDSS-like spectra to determine galactic star formation rate (SFR) over cosmic time. Wu & Boada (2019) noted that increasing the number of photometric filter bands used to train the neural network improved the accuracy of the predicted metallicity value of the galaxy. Therefore, the application of neural networks to narrow band photometric surveys, as in this paper, is an obvious step in deriving galactic evolution parameters. This paper is a proof of concept study, investigating whether neural networks can be used to derive the age and metallicity parameters from narrow-band-photometric data. We also examine how the accuracy of recovering age and metallicity gradients, compared to those derived directly from the spectra, depends on the training set use in the neural network.

In the next section, the synthesis of the data is discussed. This is followed by the methodology of the neural network and analysing gradient retrieval in Section 3. Section 4 presents the results of the neural network. Discussion and conclusions are provided in Section 5.

2 DATA

In this paper, we develop a neural network model to derive metallicity from the narrow-band filter photometric data, similar to the data that will potentially be gathered by the J-PAS survey. We targeted the J-PAS survey because it is the next-generation large-scale survey, and a computationally efficient analysis tool is required to derive stellar population properties for the many pixels of photometric data. To this end, we construct J-PAS-like narrow band filter data, i.e. ‘mock J-PAS data’, from CALIFA IFU spectra. We then assume that the spectroscopically derived ages and metallicities from the CALIFA data are the true values for each spectrum within each galaxy. The training and testing data sets for our neural network are composed of the mock J-PAS data and the spectroscopically derived age and metallicity. In Section 2.1, we explain the CALIFA data, and in Section 2.2 we describe how we make the synthesized J-PAS data from the CALIFA spectra.

2.1 CALIFA

The CALIFA survey (Sánchez et al. 2012) used the PMAS/ PPAK integral field spectrograph, mounted on the Calar Alto 3.5 m telescope. Each galaxy in the data set was observed three times, with dithering used to reach a spectral resolution of ∼1 arcsec. The integral field unit (IFU) allows 2D spectra in a grid over the surface of the galaxy to be collected, through exposure times of 1800 and 900 s for the blue and red gratings, respectively. The CALIFA parent sample consists of 937 galaxies selected from SDSS DR7 within $0.005 < z < 0.03$, with the majority being field galaxies. From the parent sample, ∼600 galaxies were observed with a diameter limit to fit within the IFU field of view and down to $M_{\rm B} \sim -18.0$ mag (for more information about the CALIFA sample, see Sánchez et al. 2012; Walcher et al. 2014).

Galactic spectra are binned, and the code GANDALF (Sarzi et al. 2006) is applied to them. GANDALF simultaneously fits the absorption and emission lines, treating the latter as additional Gaussians. In the first step, emission lines are masked and the absorption line spectrum is fitted as the penalized pixel-fitting PPXF (Cappellari & Emsellem 2004), using the stellar population models of Vazdekis et al. (2010) as templates. In this step, radial velocities and absorption line broadening for the stellar components were derived. The best values of velocity and broadening and the best template mix are then used as initial values for the calculation of emission lines. The fit allows for low order Legendre polynomials in order to account for small differences in the continuum shape between the pixel spectra and the templates. Emission lines were subtracted from the observed spectra before extracting their star formation histories.

Star formation histories were derived using the code STEllar Content and Kinematics via Maximum A Posteriori likelihood (STECKMAP; Ocvirk et al. 2006) on the emission line-cleaned spectra as described above, with the MILES stellar library (Sánchez-Blázquez et al. 2006), a Kroupa Universal initial mass function (Kroupa 2001) and Padova 2000 (Girardi et al. 2000) isochrones, which cover a range of ages and metallicities from 63 Myr to 17.8 Gyr and $-2.32 <$ [Z/H] $< +0.2$, respectively (for a detailed description of the procedure see Sánchez-Blázquez et al. 2014). No cosmological priors were applied when the values for the ages of the stellar populations were determined. This means that the ages of the galaxies are allowed to be, in principle, higher than the age of the Universe. In a number of cases, we also run STECKMAP and mask the position of the emission lines instead of subtracting them, obtaining the same results (the differences in the mean values of ages and metallicities is lower than the random errors due to the noise of the spectra).

We have decided to use IFU data as it is the most suitable for radial gradient analysis of galaxies. IFU data allow better spatial averaging of galactic properties than long-slit instruments. The sample used in this analysis, taken from Sánchez-Blázquez et al. (2014), comprises a total of 190 galaxies with high enough quality data to compute age and metallicity. Of this sample, 44 galaxies are early-type galaxies and 146 are late types according to their classification on the SIMBAD data base (Wenger et al. 2000). This is not representative of the full CALIFA sample (Walcher et al. 2014) that contains a significantly higher fraction of elliptical galaxies. From the star formation history and age – metallicity relation derived with STECKMAP, we calculate a mean luminosity-weighted age and metallicity for each spectrum in the data set using spectral fitting. Any spectra whose fit was deemed to be poor (i.e. with reduced $\chi^2 > 2$) were ignored for this work, giving a data set composed of 19 727 spectra.

2.2 Synthesized J-PAS data

The J-PAS survey is a multiband photometric survey that runs at the Observatoro Astrofisico de Javalambre in Spain, with a $3.89m^2$ collecting mirror. The J-PAS instrument covers a 4.7 deg^2 per observation, with a pixel size of 0.456 arcsec. The effective integration time is 4.96 h per field (Benitez et al. 2014).

The response curve of the 54 narrow-band filters are spaced 100 Å apart with a full width at half-maximum of 145 Å, covering the range of 3785−9100 Å. The magnitude limit is $21.0 < m_{AB} < 25.7$ mag, and varies by filters. These narrow-band filters act as a low-resolution spectrograph, with an effective resolution of 100 Å (compared to CALIFA's resolution of 2Å) and are able to detect the broad galaxy emission features. A comparison of the CALIFA spectrum and synthesised J-PAS SED is shown in Fig. 1 for a spectrum taken from NGC 2530. It can be seen that there is a loss of all spectral line information due to the decreased resolution which increases the difficulty of determining stellar populations properties.

The synthetic photometry was obtained by convolving each spectrum with the response function of the J-PAS filters. As the spectral range of CALIFA is 3700–7000 Å, we only measured 36 J-PAS magnitudes. We further exclude the two bands JPAS-6600 and JPAS-6700 to avoid being affected by the possible presence of the H α emission line.

The determination of ages and metallicities using broad-band colours is difficult due to the similar variations in the shape of the continuum caused by an increase of both parameters (the so-called age–metallicity degeneracy, Worthey 1994). Individual absorption lines are also affected by this problem, but each line has a different sensitivity to variations of age versus metallicity and therefore if we measure several lines we can alleviate the problem.

However, the usefulness of narrow-band photometry to derive stellar population properties has not been sufficiently explored. These magnitudes are much more sensitive to the strength of absorption lines than broader bands and they can be measured with a much larger signal-to-noise ratio than the absorption lines. A derivation of age and metallicity using medium-band photometry from the ALHAMBRA survey was presented in Díaz-García et al. (2015). This study showed that age and metallicity can be measured with an rms uncertainty of 0.10 and 0.16 dex, respectively.

The increased number of J-PAS bands, compared to those of ALHAMBRA, mean that we have more information available to circumvent large errors caused by the age–metallicity degeneracy. Using SED fitting to the SDSS spectroscopy data and the J-PAS mock data created from the spectroscopy data, Mejía-Narváez et al. (2017) demonstrated that the age and metallicities of the galaxies can be obtained from the J-PAS data as accurately as from the spectroscopy data. Our work is motivated by their study showing promising results that J-PAS-like narrow-band data contain some information to break the age and metallicity degeneracies in a similar degree to the spectroscopy data. Hence, it would be interesting to explore if the neural network can learn such information and provide the accurate age and metallicity much faster than traditional methods.

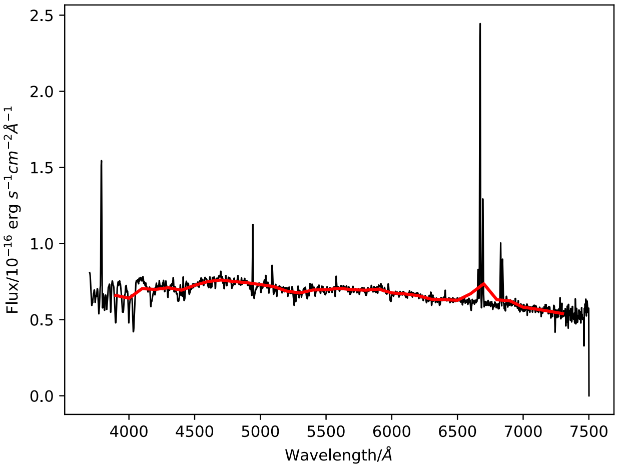


Figure 1. A comparison of the spectral curve given by CALIFA (black) and the simulated J-PAS response (red) for one spectrum in NGC 2530. The majority of spectral lines cannot be seen in the J-PAS SED, making it more difficult to extract age and metallicity information.

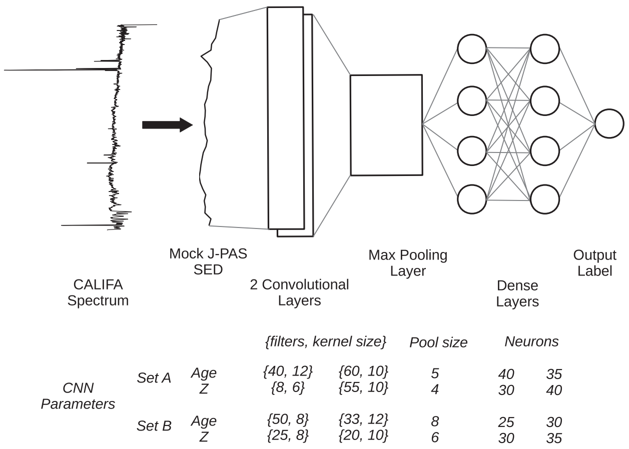
3 METHOD

3.1 Neural network

We use supervised neural networks to predict the metallicity and age of a sample of galaxies from their J-PAS-like SEDs (see Section 2.2 for details on their synthesis) with the TENSORFLOW KERAS API (Abadi et al. 2015).[1] The determination of age and metallicity are treated as a regression problem. The convolutional neural network (CNN) we develop uses the spectroscopic age and metallicity derived by CALIFA as the 'true' value for the purposes of training. Each of the neurons in the network begins with some randomized weight, and the simulated magnitudes for each band pass through the CNN to calculate a predicted value for the age or metallicity. The mean-squared error (MSE) of predicted versus spectroscopic age or metallicity is back propagated through the network to adjust the weights of the neurons. This process is repeated to obtain an accurate output.

The CNN used in this work has an architecture as illustrated in Fig. 2. The starting point for the CNN was taken from Fabbro et al. (2018), who used a CNN to analyse stellar spectra. Our chosen architecture has two convolutional layers, a max pooling layer and two dense layers. The 1D convolutional layers capture patterns and multifilter features across the SED. The max pooling layer then reduces the dimensions of the convolutional layers' output. This is applied to the classical dense neural network layers that calculate the age or metallicity via non-linear combinations of values given by the outputs of the max pooling layer. We experimented with

[1] See https://github.com/ChoongLing/SimulatedJ-PAS for the code used for the methods discussed in this section.



			{filters, kernel size}		Pool size	Neurons	
CNN Parameters	Set A	Age	{40, 12}	{60, 10}	5	40	35
		Z	{8, 6}	{55, 10}	4	30	40
	Set B	Age	{50, 8}	{33, 12}	8	25	30
		Z	{25, 8}	{20, 10}	6	30	35

Figure 2. A schematic view of the architecture used for the convolutional neural network (CNN). The CALIFA spectra are converted into mock J-PAS photo-SED, which are then passed through two convolutional layers. A max pooling layer reduces dimensionality, and its results are passed through a single dense layer. The predicted value of age or metallicity is then output by the CNN. The hyperparameters used for our CNNs are shown at the bottom of this figure.

architectures containing 1, 2, and 3 dense layers and found that models with two dense layers provide the most accurate predictions for both the age and metallicity of our data. The age and metallicity for each of Sets A and B (see Section 3.2) were determined by separate CNN models, which had identical architectures but different hyperparameters, which are shown at the bottom of Fig. 2. The layers' hyperparameters were optimised by Hyperas.[2] Comparisons showed that the set of hyperparameters chosen by Hyperas provide more accurate predictions than are made by CNNs with manually chosen hyperparameters.

We also adopted early stopping with a patience parameter of 25 for the CNN. This meant that if there was no improvement in the MSE of the parameter recovery after 25 epochs, training would stop. The CNN would train for a maximum of 5000 epochs or until the MSE stabilised. A total of 19 727 spectra from 190 of galaxies was used in this analysis.

To train the neural network to predict metallicity and age for the full data set, 25 per cent of the data was kept aside for the testing of the trained CNN to produce our results. The other three quarters was used for training the CNN. This process was repeated three more times so that metallicity and age predictions were made for the full data set, with each iteration using a training set independent of the unseen testing set. Our final results are given by single realizations of the trained CNN models. The randomness in predictions is taken into account when gradients are calculated (see Sections 4.2 and 4.3) but not for individual predictions. This is because we do not have values for the errors of our spectroscopically determined ages and metallicities, and therefore we could not properly estimate the uncertainties of the model prediction, for example, through using Bayesian neural networks (e.g. Ciucă et al. 2020). Additionally, our CNNs are not designed or trained to handle noise. Although evaluating the uncertainties of individual predictions is important, it is beyond the scope of this study because the aim of this proof of concept study is to demonstrate the ability of CNNs to extract age and metallicity data from narrow-band spectra.

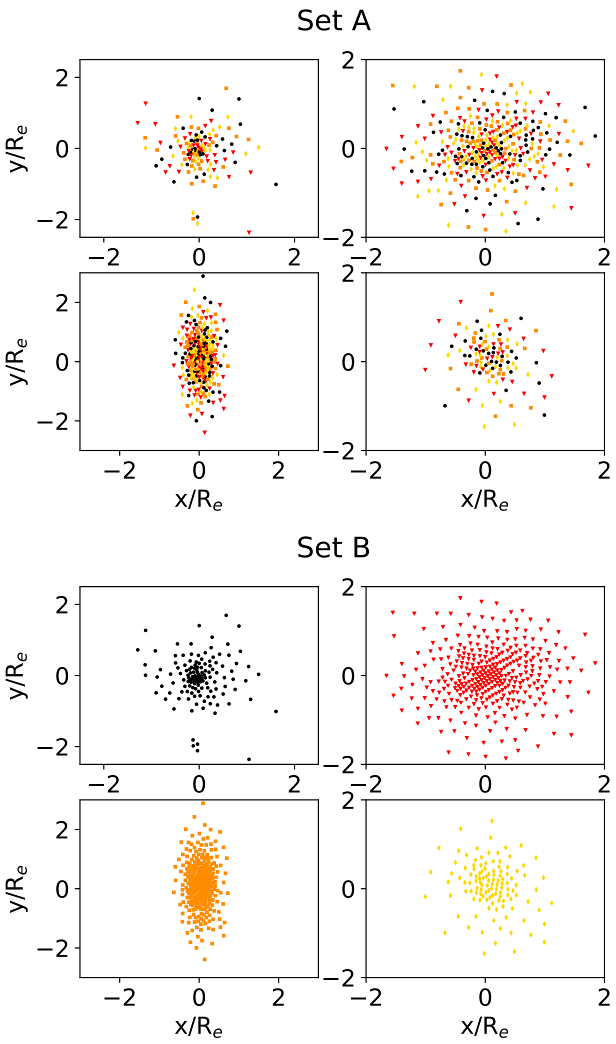


Figure 3. Illustrations showing how the spectral data are split into four subsets, as described in Section 3.2. The top four panels show the splitting for Set A and the lower four for Set B. In both sets of panels, the spatial distribution of the spectra in four different galaxies are shown. Each spectrum is represented by a coloured shape depending on which subset it belongs to (the black circles, the red triangles, the orange squares, or the yellow diamonds). In Set A, the spectra within each galaxy are split amongst the four subsets, whereas in Set B all of the spectra for a given galaxy are in the same subset.

3.2 Defining the training and testing sets

Two ways of splitting the data set into four subsets are explored in this work, which are illustrated in Fig. 3. The first is by splitting the spectra within each galaxy randomly into the four subsets, ensuring that one quarter of the data from each galaxy are put into each one of the four subsets. The CNN is then trained on three of the four subsets, with the final subset kept aside and unseen for testing. This will be referred to as Set A. The other method, Set B, is created by randomly splitting the 190 galaxies into four subsets, with all of the spectra from one galaxy in the same subset. This means that the testing set for Set B contains galaxies that have not been seen at all by the CNN during training. The key difference is that in Set A the training set contains spectral data from every galaxy therefore the training and testing data sets are not completely independent due to the covariance between adjacent spectra.

[2] https://github.com/maxpumperla/hyperas

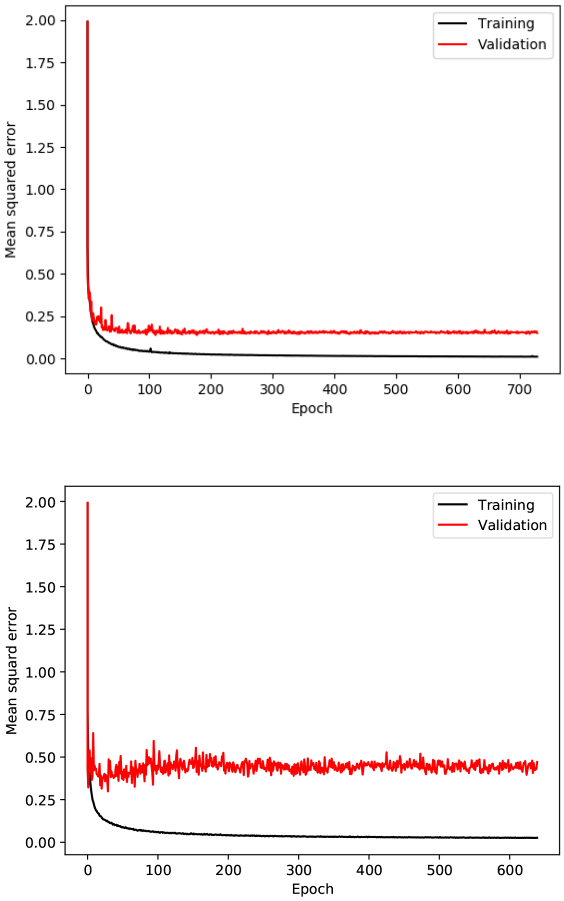


Figure 4. Learning curves for the training of one CNN model for the ages of galaxies in Set A (upper) and Set B (lower). The black curve shows the mean-squared error (MSE) of the training set, and the red curve shows the MSE of the validation set. The initial MSE for the 0th epoch is set as 0.199, which is calculated from a set of random predictions. Early stopping with a patience parameter of 25 was used when training the CNNs.

It is possible that spectra from the same galaxy will have similar stellar and chemical evolution histories, even at different positions within the galaxy. In this way, Set A mimics a situation where a large number of galaxies are included in the training set, which will cover the diversity in galactic evolutionary history, so that the training set contains data from similar galaxies to those in the application set. Set B demonstrates the realistic case, where we do not have any previous knowledge about a galaxy in the testing set. In this proof of concept study, we compare the ideal case of Set A with the realistic case in Set B. Although it is more realistic, Set B suffers due to the relatively small size of our data set. Conversely, Set A is a suitable way of exploring the potential benefits of a large, comprehensive training data set. Therefore, this comparison will show the potential of the CNN method when a large data set becomes available in the future.

The learning curves obtained when training the CNNs to predict ages are presented in Fig. 4, using MSE as the metric. The MSE for the 0th epoch is set at 0.199, which is the MSE of random predictions made from the set of age labels. It can be seen that the MSE of the training sets converges quickly, but the validation set (which is not used for training, and is also the set that the final trained CNN model will be applied to) does not reach the same level of accuracy as the training set. In the idealized case of Set A, there is a small difference between the MSE of the training and validation data. This implies that there is a subtle difference between the properties of galactic spectra at different regions within the galaxy as the training and validation data came from different spaxels, but the same galaxy.

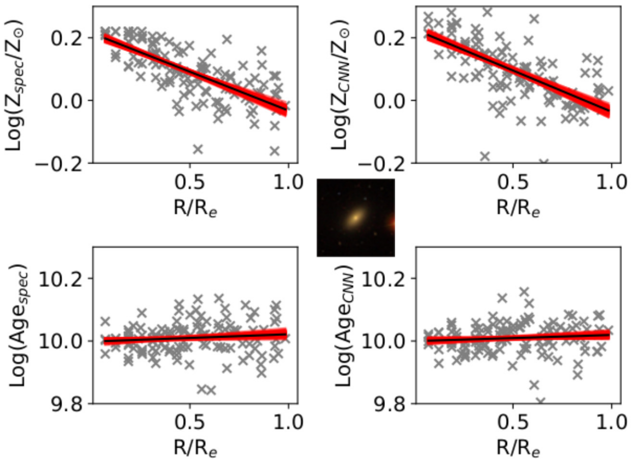


Figure 5. The derived spectroscopic and CNN-predicted ages and metallicities against radius for NGC 7671 with a 4 arcmin × 4 arcmin SDSS image embedded in the centre. The CNN trained using Set A (see Section 3.2). The top row shows metallicity and the bottom panels display age. The left-hand column shows the parameter values derived directly from CALIFA spectra, and the right contains predictions from the CNN. The value of each spectrum is shown as the grey crosses. The linear fits to these data computed by 100 iterations of MC bootstrapping are shown as the red lines, with the mean values for these fits plotted as the solid black line.

3.3 Radial gradient analysis

Radial gradients for the age and metallicity within the effective (half-light) radius, R_e of the galaxy are also calculated and analysed for both the CNN predictions (Section 3.1) and CALIFA spectroscopic age and metallicity. We analysed the gradients only for the galaxies that have at least 25 spectral data points within $R < R_e$ and there is at least one data point at $R > R_e$, to ensure that enough spectra to cover up to $R < R_e$. This allows us to produce reliable radial gradients.

To obtain the gradient, the inclination of each galaxy was corrected to determine the face-on projected radius for the position of each spectrum. A linear fit to age or metallicity against radius was computed using Monte Carlo (MC) bootstrapping to randomly select a sample of 75 per cent of the data. A least-squares fit was obtained for 100 MC samples. Then, the mean gradient and its standard deviation were calculated from these samples. This was performed on both the spectroscopic and CNN predicted values, which were then compared. As no uncertainties were computed from the CNN predictions or spectroscopic values, the uncertainty in the gradient fitting was determined by the standard deviation of the MC derived gradients. Therefore, the uncertainties in the linear gradient fitting do not consider any intrinsic uncertainties in the CALIFA spectroscopic analysis or CNN predictions. Fig. 5 shows an example where metallicity and age are plotted against radius for the galaxy NGC 7671 using Set A (see Section 3.2). The top row shows the spectroscopic (i.e. the true label, left) and CNN (predictions, right) metallicity, with the bottom row showing the equivalent diagrams for age. The grey crosses are the values for each spectrum. The red lines show the fits produced by each iteration of the MC bootstrapping. The black line shows the gradient derived from the mean value of the MC fits. The results of gradient analysis will be discussed in

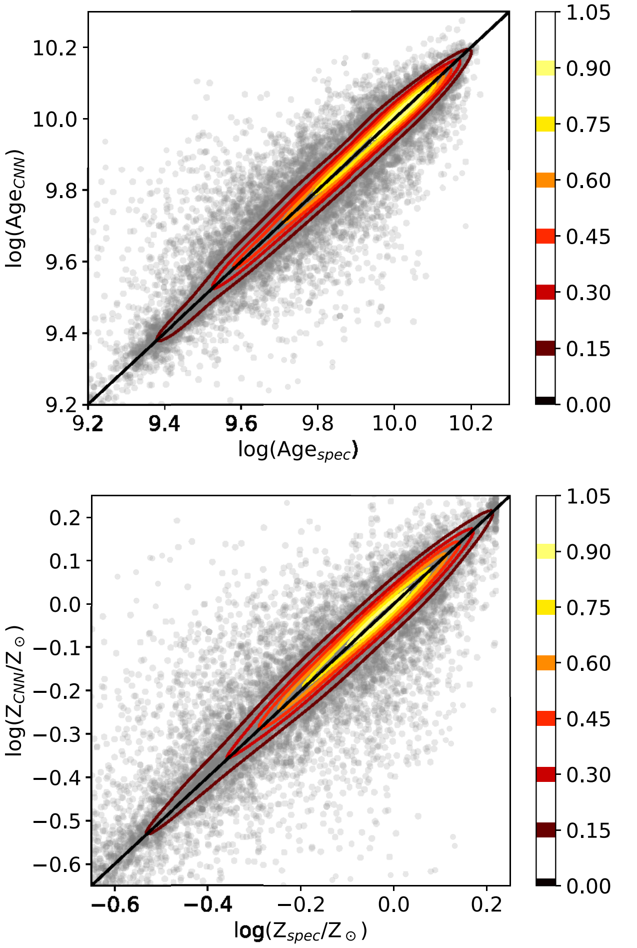


Figure 6. The luminosity-weighted age (top, Age$_{\rm CNN}$) and metallicity (top, Z$_{\rm CNN}$) derived from the CNN against the spectroscopically determined age (Age$_{q}spec$) and metallicity (Z$_{\rm spec}$) for Set A showing only data with a spectroscopically determined value of age and metallicity with reduced $\chi^2 <$ 2. The solid black line shows a 1:1 correlation, which corresponds to perfect recovery. The contour map shows the normalized density distributions of the results of the spectra. The CNN values of age and metallicity are consistent with the spectroscopically determined values, with a robust standard deviation of 0.03 dex for both values.

Sections 4.2 and 4.3. Only the gradients will be discussed in this paper.

4 RESULTS

Results from Set A will be discussed in Sections 4.1 and 4.2 and results from Set B will be presented in Section 4.3. We investigate the effects of other galactic parameters and training set size on the accuracy of CNN predictions in Sections 4.4 and 4.5, respectively. Section 4.6 covers the dependence of our radial gradients on stellar mass.

4.1 Set A: predictions of age and metallicity

The recovery of age and metallicity using Set A is shown in Fig. 6. The grey points show the prediction of the CNN against the value determined from CALIFA, which we consider to be the true values. A contour map shows the normalized distribution of these points. The solid black line shows a 1:1 correlation, i.e. a CNN prediction that is identical to the spectroscopic value. The recovery here is excellent, which can be seen as most points lie close to the 1:1 recovery line. The robust standard deviation (calculated from the median absolute deviation) of the difference between CNN and spectroscopic values for Set A are 0.03 for both age and metallicity. These uncertainties are epistemic, i.e the difference between the spectroscopic and predicted values, showing how well our CNN model is able to reproduce the spectroscopically determined values from the given set of synthetic fluxes (see Hüllermeier & Waegeman 2019, for more information). Therefore, the fact that our uncertainties here are lower than the statistical uncertainties reported elsewhere (e.g. Sánchez-Blázquez et al. 2014) is not concerning as these errors measure different effects.

This level of accuracy in reproducing age and metallicity is encouraging, and shows that the CNN is working well. Once the model has been trained, its application to the test data set is very rapid, meaning it is suitable for use in the large data sets, such as those that will be produced by J-PAS. The standard deviation in the CNN predictions is comparable to those obtained by CALIFA spectral fitting (e.g. Sánchez-Blázquez et al. 2014). The value of the Pearson's correlation coefficient (PCC) between the age and metallicity residuals of the CNN prediction and the spectroscopic values is $r = -0.24$, showing a weak negative correlation between the two predictions. This shows that our CNN models make predictions that are no more affected by the age–metallicity degeneracy than the values obtained with a full spectral fitting.

4.2 Set A: gradient analysis

The values of age and metallicity from each point – both spectral and CNN predicted – are used to calculate a radial gradient, as described in Section 3.3. The differences between the CNN predicted and spectroscopic gradients are plotted in Fig. 7. The black crosses show the difference between the calculated gradients, with the red lines showing 1σ error bars computed using the MC bootstrap sampling. The top and right-hand panels show histograms of the difference between the gradients of metallicity and age, respectively, with bins of 0.05 dex/$R_{\rm e}$. There is strong clustering of the differences in gradient in the central 0.1 dex/$R_{\rm e}$. The gradient recovery is found to be accurate to within a robust standard deviation of 0.02 dex/$R_{\rm e}$ for both age and metallicity. It can also be seen that there is no clear correlation between the age and metallicity gradient deviations of the CNN values from the spectroscopic gradients, which shows that the quality of CNN predictions are not affected by the age–metallicity degeneracy.

4.3 Set B: age and metallicity prediction and gradient analysis

The recovery of age and metallicity for Set B is shown in Fig. 8. The contour levels are the same as in Fig. 6. The epistemic robust standard deviations in this case are 0.16 dex for both age and metallicity with a PCC of the residuals of $r = -0.24$. This is the same value as the PCC in Set A, and therefore the more independent data used to train our Set B models does not affect the ability of the CNNs to overcome the age–metallicity degeneracy. It can be seen that the contours are much more spread out, and not concentrated around the black 1:1 recovery line. The age recovery, in particular, shows an offset with CNN predictions systematically lower than the spectroscopic values. At lower metallicities, the predictions of the CNN become less accurate, which can be seen as the contours spread further from the black 1:1 line. This effect is likely due to the rarity of spectra with log ($Z_{\rm spec}/Z_\odot$) < -0.75 in the training set. The use of synthetic spectra or data

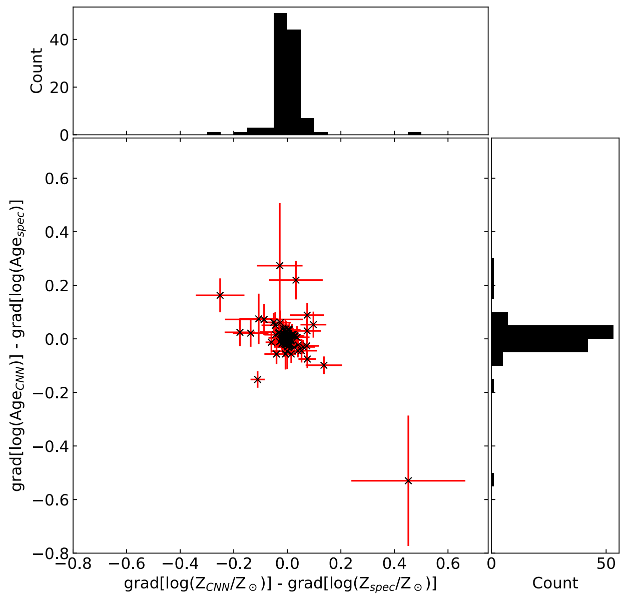


Figure 7. The difference between the gradients from CNN predicted age, grad(log(Age$_{\rm CNN}$)), and the spectroscopically derived age, grad(log(Age$_{\rm spec}$)), against the difference in the CNN predicted metallicity gradient, grad(log(Z$_{\rm CNN}$/Z$_\odot$)), and spectroscopically derived metallicity, grad(log(Z$_{\rm spec}$/Z$_\odot$)). The red error bars show 1σ confidence limits for the gradient fitting. The top and right-hand panels show histograms of the gradient differences in bins of 0.05 dex/$R_{\rm e}$. The robust standard deviation for the difference in gradients is 0.02 dex/$R_{\rm e}$ for both age and metallicity. There is no visible correlation between differences in CNN predictions for age and metallicity gradient and the respective spectroscopic gradients.

augmentation (e.g. Ciucă et al. 2020) could improve predictions by creating more training examples for lower metallicity data points and will be considered in future works.

The quality of the CNN's gradient recovery of the spectroscopic values in Set B are displayed in Fig. 9. These are markedly worse than the results obtained in Set A. In this case, the standard deviation for gradient recovery, grad$_{\rm CNN}-$grad$_{\rm spec}$, is 0.15 dex/$R_{\rm e}$ and 0.16 dex/$R_{\rm e}$ for age and metallicity, respectively. The reason for this discrepancy between Sets A and B is likely due to the diversity in star formation histories among galaxies. The accuracy of Set A implies that the formation history of different regions within the galaxy are similar. As a result, the training set of Set A contains data with similar stellar populations to the testing set, which improved the performance of the CNN. Conversely, the training set for Set B does not contain enough variation to cover the star formation and chemical evolution histories of the unseen galaxies for the CNN to accurately reproduce the spectroscopic values of age and metallicity. This could be resolved in future works by either using a larger data set or employing synthetic data to increase the diversity of our training set.

4.4 Dependence of predictions on galactic parameters

To study the importance of the similarity of stellar populations between the training and testing sets, we explore the dependence of the accuracy of CNN predictions of age and metallicity on specific SFR (SFR/M_*), i.e. the total galactic SFR divided by its stellar mass.), galactic inclination, extinction ($A_{\rm V}$), and galaxy morphology. We also examined the effect of the fractional size of the galactic bulge on the accuracy of our predictions but found that there was no visible correlation.

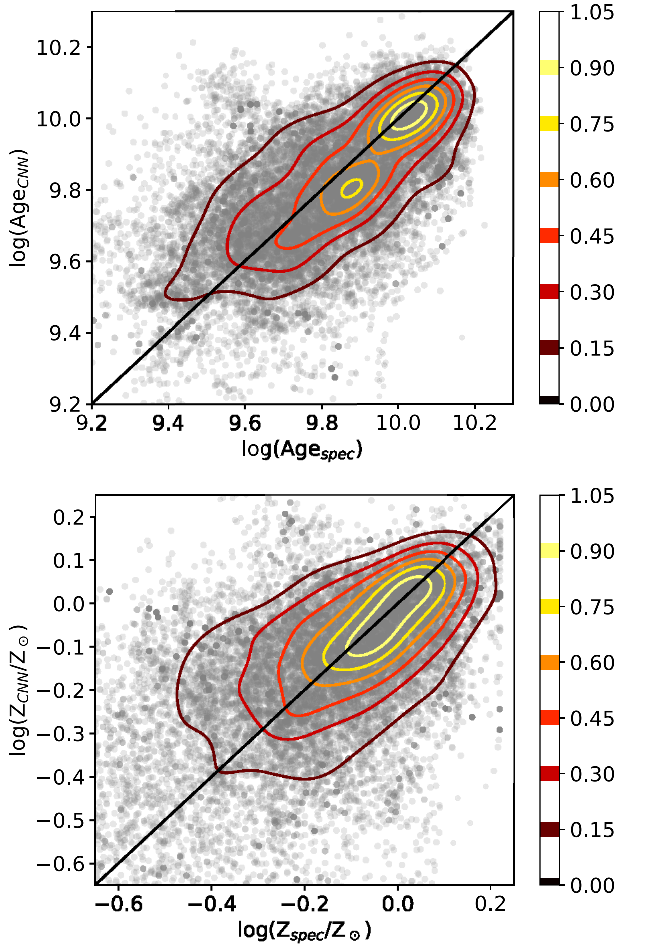


Figure 8. The luminosity-weighted age (Age$_{\rm CNN}$, upper panel) and metallicity (Z$_{\rm CNN}$, lower panel) derived from the CNN against the spectroscopically determined age (Age$_{\rm spec}$) and metallicity (Z$_{\rm spec}$) for Set B. Recovery here is significantly worse than in Set A, with robust standard deviation of 0.14 and 0.16 dex for age and metallicity, respectively.

The median and robust standard deviation of the difference between the CNN predictions and spectroscopically derived values are computed from each SED within the galaxy. In Figs 10, 11, and 12, the median values for each galaxy are shown in the left-hand column of the figures. The robust standard deviations for each galaxy are given in the right-hand columns. Each of these values are plotted against specific SFR, inclination and $A_{\rm V}$ in Figs 10, 11, and 12, respectively. The upper four panels in each figure show the results of Set A, while the lower four panels show the results of Set B. The first and third row of panels for each set corresponds to the metallicity and the second and fourth rows correspond to age.

In Fig. 10, it can be seen that the robust standard deviation for accuracy of predictions of age and metallicity slightly increases with specific SFR. However, this is not reflected in the median values. Additionally, there is no visible trend in either median or robust standard deviation of predictions with inclination (Fig. 11) or extinction (Fig. 12), which shows that our CNN models are not affected severely by these galactic properties.

The 190 galaxies in our sample were split by morphology (taken from the SIMBAD data base, Wenger et al. 2000) giving 44 early-type galaxies and 146 late-type galaxies. CNNs were trained on 33 of the elliptical galaxies and 114 spiral galaxies, respectively, using the

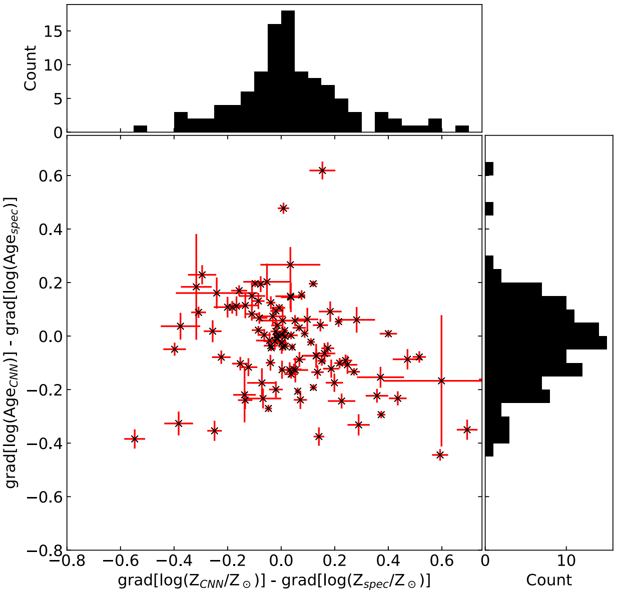


Figure 9. The difference between the gradients from CNN predicted age, grad(log(Age$_{\rm CNN}$)), and the spectroscopically derived age, grad(log(Age$_{\rm spec}$)), against the difference in metallicity gradient from the CNN, grad(log(Z$_{\rm CNN}$/Z$_\odot$)), and spectroscopically derived metallicity, grad(log(Z$_{\rm spec}$/Z$_\odot$)) for Set B. The recovery in Set B is much worse than Set A, with robust standard deviation increased to 0.15 dex/$R_{\rm e}$ and 0.16 dex/$R_{\rm e}$ for age and metallicity, respectively.

method for Set B, as in Section 4.3. These CNNs were then applied separately to the remaining galaxies in each morphology set.

The robust standard deviations for the differences between spectroscopic and CNN predicted values are given in Table 1. It can be seen that predictions for the ages of each of the morphology groups are more accurate when the CNN has been trained on the same morphology group. Additionally, when the CNN has been trained on only early-type galaxies, the age prediction performs best for early-type galaxies and has a robust standard deviation of 0.10 dex. Predictions of the age and metallicity of late-type galaxies are of similar quality regardless of whether the CNN is trained on early- or late-type galaxies. This is unexpected, but is likely due to the presence of similar stellar populations between early-type galaxies and the bulges of late-type galaxies. Overall, the recovery of early type galactic properties is significantly better than the full data set for Set B, whose values are shown in the third row and column of each table, but is still worse than for Set A. We believe that the increased accuracy in recovery of early-type galaxies is due to the greater degree of similarity between the stellar populations found in early types than between late types. This supports our conclusion that the CNN is more capable of predicting age and metallicity values for stellar populations similar to those present in the training set. Therefore, a larger, high-quality data set would be crucial for future deep learning analysis of stellar populations.

4.5 Training set size

The size of the training set is very important in neural networks. Typically, very large data sets are used in analysis using CNNs. This is because a large volume of data increases the accuracy of neural network predictions. In this section, we discuss the impact of how the size of the training set affects the predictions of our CNN model, though we are still limited by our relatively small data set.

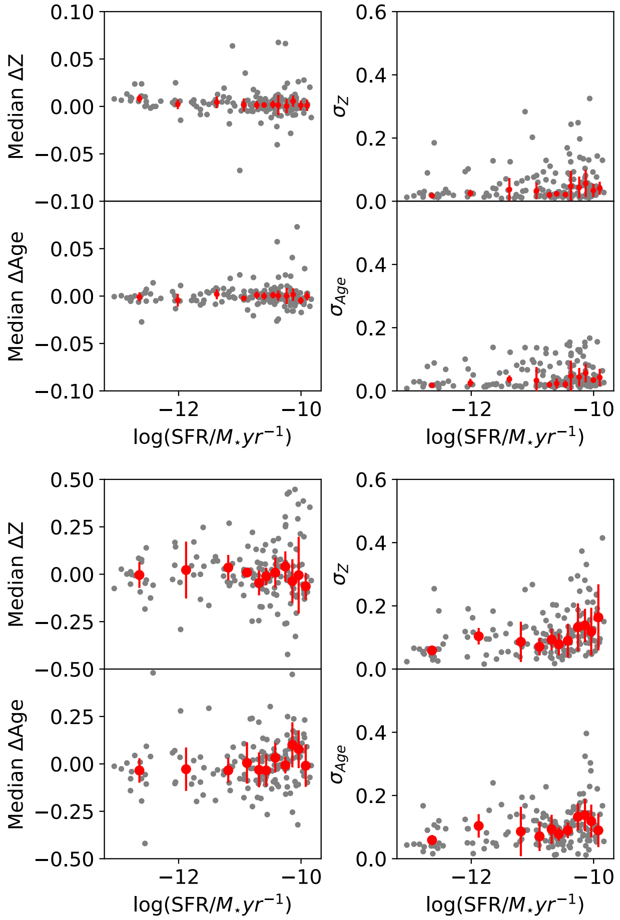


Figure 10. The dependence of the accuracy of predictions on specific star formation rate (SFR) for Set A (upper four panels) and Set B (lower four panels). The left-hand column shows the median value of the difference between CNN predictions and spectroscopic values of age (median[log ($Age_{\rm CNN}$) − log ($Age_{\rm spec}$)], 2nd and bottom row) and metallicity (median[log ($Z_{\rm CNN}/Z_\odot$) − log ($Z_{\rm spec}/Z_\odot$)], first and third rows). The right-hand column shows the robust standard deviation of the difference between CNN predictions and spectroscopically determined values. The grey dots show the median and robust standard deviation computed from the results of different SEDs for each galaxy's age or metallicity against specific SFR. The red dots show the median of bins of 16 galaxies and the error bars show the robust standard deviation of the bin. It can be seen that the uncertainty of the predictions increases slightly with specific SFR, though the median values do not show such a dependence. Note the y-axis for the median differences of Set A has been reduced by a factor of 5 due to the significant difference in accuracy.

Fig. 13 shows the robust standard deviation of the difference between spectroscopic and CNN predicted age values for Set A (the solid lines) and Set B (the dashed lines) as a function of the training set size, given as a fraction of the total size of the data set. Note that we only used the results for data points whose spectroscopic values are reliable (i.e. with reduced $\chi^2 < 2$), to evaluate the performance when the CNN model is applied to the similar quality data to the training set. Training and application of the CNN model was performed 100 times with randomly selected training and application sets for each iteration. The standard deviation for the recovery of age was recorded for each model, and the mean and uncertainty of these standard deviations is shown in Fig. 13. The horizontal red-dotted

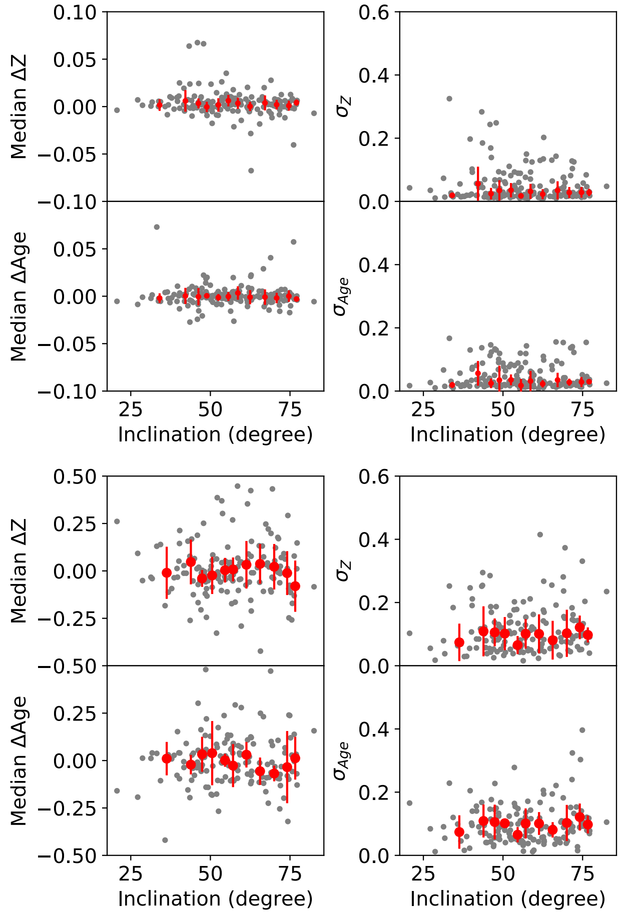


Figure 11. As in Fig. 10 but plotted against inclination. The inclination has been adjusted to be between 0 and 90°.

line shows the standard deviation we would expect if the predictions were made by simply choosing a random value from the set of spectroscopic ages. Both Set A and Set B results are below this line, which confirms that the CNN learned some relation to map the input features to the output values better than picking a random value from the training set.

It can be seen that the accuracy of recovery of Set B decreases as the training set size decreases, and the uncertainty of this accuracy increases. For Set A, the decrease in the accuracy of recovery between 5 per cent and 75 per cent is $\sim 1\sigma$ so is not statistically significant. Despite the increase in prediction accuracy for Set B, the recovery in Set A with a training set of 5 per cent of the total data set is ~0.07 dex smaller than the recovery of ages in Set B using 75 per cent of the data set. This supports our conclusion that increasing the number of galaxies in our data set to account for the diversity in star formation histories is crucial in increasing the accuracy of CNN predictions. In other words, the number and diversity of the spectroscopic data used in this paper is not enough for accurate recovery of stellar population parameters from a testing set composed of galaxies that are not included in the training set. We would expect that with data from more galaxies with a diverse range of star formation histories, either real or simulated, we would see the prediction accuracy for Set A to improve with increasing sample set size, as seen in Set B in Fig. 13. In addition, we expect that the accuracy of the recovery for Set B, when using a large training set, would approach that of Set A.

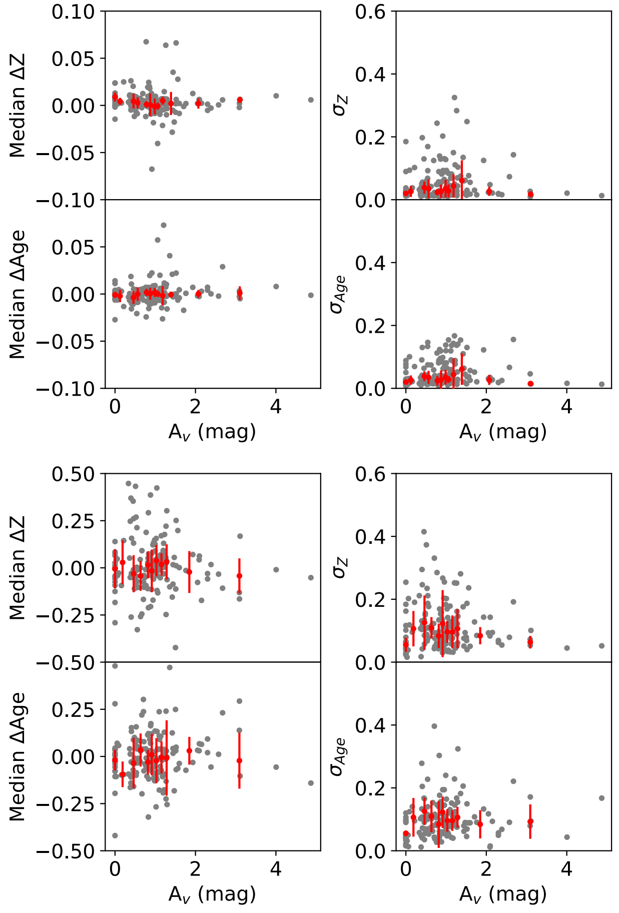


Figure 12. As in Fig. 10 but plotted against extinction, A_V. There does not appear to be any correlation between the accuracy of predictions and the extinction.

Table 1. The robust standard deviations of the difference between spectroscopic and CNN predicted age (upper) and metallicity (lower table) defining the training and application sets as for Set B (see Section 4.3). The columns indicate whether the CNN was trained on early or late-type galaxies, and the rows indicate whether the application set (appl set) set was composed of early-type or late-type galaxies. The uncertainty for the full set, as derived in Section 4.3, is given in the third row and column of each table for comparison. See the text for more information.

Age		Training set		
		Early types	Late-types	Full set
Appl	Early types	0.10	0.14	
Set	Late-types	0.16	0.15	
	Full set			0.14
Z		Training Set		
		Early types	Late-types	Full Set
Appl	Early types	0.12	0.13	
Set	Late-types	0.19	0.17	
	Full set			0.16

These findings imply that the stellar populations in different regions within the same galaxy are significantly more similar than stellar populations in different galaxies with the same age and metallicity. Therefore, in order to use CNNs to predict the age and metallicity in a galaxy, we require a very large training data set, covering the full parameter space of stellar population properties.

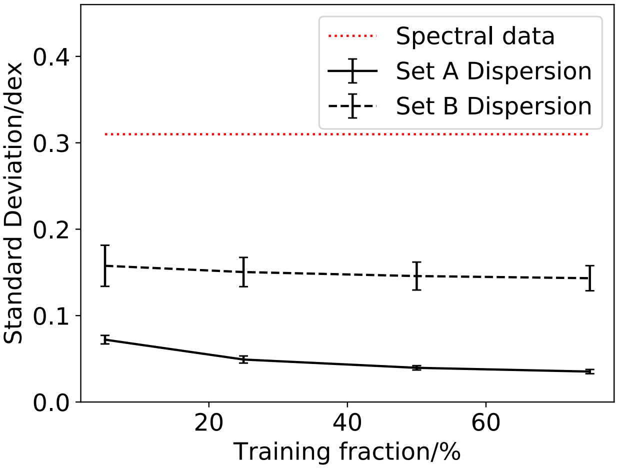


Figure 13. The variation in standard deviation of CNN recovery of age values as a function of the size of the training set is varied. The training set varied between 14795 (75 per cent of the full data set) and 986 (5 per cent) SEDs for Set A (solid line). Training with Set B uses between 157 galaxies (75 per cent) and 10 galaxies (5 per cent; the dashed line). The red-dotted line shows the standard deviation we would expect if the predictions were made by simply choosing a random value from the set of spectroscopic ages.

4.6 Mass dependence of radial gradients

The dependence of age gradients on galactic stellar mass is of interest when evaluating how galaxies evolve. The relationships we have found between these quantities are shown in Fig. 14. The left-hand panel in this figure shows the gradients derived from the spectroscopically measured age. The relationship of the late-type (the black squares) galaxies' age gradients on mass resembles that of Fig. 6 from Sánchez-Blázquez et al. (2014), which uses the same spectroscopically derived age values as this paper. This demonstrates that our method of gradient derivation provides consistent results to that of the previous work. It can be seen that the gradients produced by our analysis from Set A (central panel) is similar to that of the gradients derived from spectral values (left-hand panel) and therefore showing similar trends to Sánchez-Blázquez et al. (2014). Conversely, Set B (right-hand panel) shows significant differences from the gradients calculated from the spectroscopically derived age (left-hand panel), which can be seen in both the medians for stellar mass bins (the filled symbols) and the derived gradient for individual galaxies (the open symbols).

The mass dependence of age gradients for a variety of galactic morphologies was studied in González Delgado et al. (2015). In Fig. 10 of their paper, the early-type galaxies show higher values of the age gradient in the higher mass galaxies at $\log(M_*) \gtrsim 10.5$. The late-type galaxies show similar trends in the same mass range, but show systematically lower gradient than the early-type galaxies. Then, at $\log(M_*) \lesssim 10.5$ the gradient values become larger for the smaller mass galaxies in the late-type galaxies. These trends are qualitatively reproduced in the left-hand panel of Fig. 14. However, the values of the gradients we derived here are systematically higher than those in González Delgado et al. (2015). This could be due to the differing methods of gradient derivation or differences in stellar population modelling (see González Delgado et al. 2015, for details).

5 SUMMARY AND DISCUSSION

We present a proof of concept study of an application of a CNN model to recover age and metallicity of nearby galaxies. The data used in this work are taken from the CALIFA data set and is synthesized to produce data resembling 36 J-PAS-like photometric bands that were used to train a CNN model. A total of 21 230 spectra from 190 galaxies are used in this analysis. The CNN was able to predict age and metallicity accurately in the ideal case of Set A (Sections 4.1 and 4.2), where the data used in both the training and application sets came from spectra in the same galaxies. The recovery for age and metallicity is excellent and has a robust standard deviation of 0.03 dex. The radial gradients of age and metallicity are calculated from the CALIFA spectroscopically derived age and metallicity, and the CNN predictions of these values for each galaxy. The robust standard deviation of the difference between the gradients with spectroscopically derived values and the CNN predicted values is 0.0 dex/R_e for both age and metallicity. Radial gradients are also recovered well with the CNN.

On the other hand, for the more realistic case of Set B (Section 4.3), where the training and application data sets are composed of spectra located in different galaxies, the CNN's recovery of age and metallicity is markedly worse. The robust standard deviation for the recovery in Set B is a factor of ~7 worse for age and ~8 worse for metallicity than Set A. There is also a significant degree of difference between the radial gradients derived from the spectroscopically measured values and those calculated using predictions from the CNN trained using Set B, due to the greater dispersion of CNN predictions for each spectra. We attribute this decrease in prediction accuracy with respect to Set A to the lesser degree of similarity in stellar populations between different galaxies compared to different regions within the same galaxy. This is supported by the smaller error in recovery for early-type galaxies compared to late-type galaxies in Set B, as the latter have a greater range of stellar populations. Our data set contains a relatively small number of galaxies, which was not enough to account for the vast diversity of stellar populations. If we had a larger number of galaxies with a great enough overlap of stellar properties, we expect that the CNN predictions would improve greatly and approach the level of accuracy obtained by Set A.

In this work, only the errors from gradient fitting are considered. An improvement to the method would be to consider the error in the CNN predictions of age and metallicity. This would be an important step in properly evaluating the uncertainties of the CNN predictions for the analysis of real observational data.

We have demonstrated that the CNN model is able to predict age and metallicity values on a relatively small proportion of the training set provided it has enough high-quality data to cover the range of stellar populations present in the application set. We show our models are not strongly affected by degeneracies with SFR, relative bulge size, inclination angle, or extinction. This, along with the low computing power required to apply the trained model to new data, makes CNNs a suitable method of analysis for large data sets such as those that will be produced by the J-PAS survey. However, constructing a large enough high-quality training data set to improve machine-learning models is crucial. Therefore, we will continue to need additional large spectroscopic surveys and high-performance spectral fitting codes. More high quality spectral (preferably IFU) data and sophisticated stellar population models to fit these spectra would be invaluable for creating a high quality training set for further neural network studies. The efforts in increasing the coverage of IFU surveys, such as SAMI (Croom et al. 2012) and MaNGA (Bundy et al. 2015), and their improving fitting pipelines will be essential in future applications of CNNs to situations similar to that of Set B in this work. Additionally, the use of synthetic spectra from simulated galaxies with a large range of evolutionary histories could also be

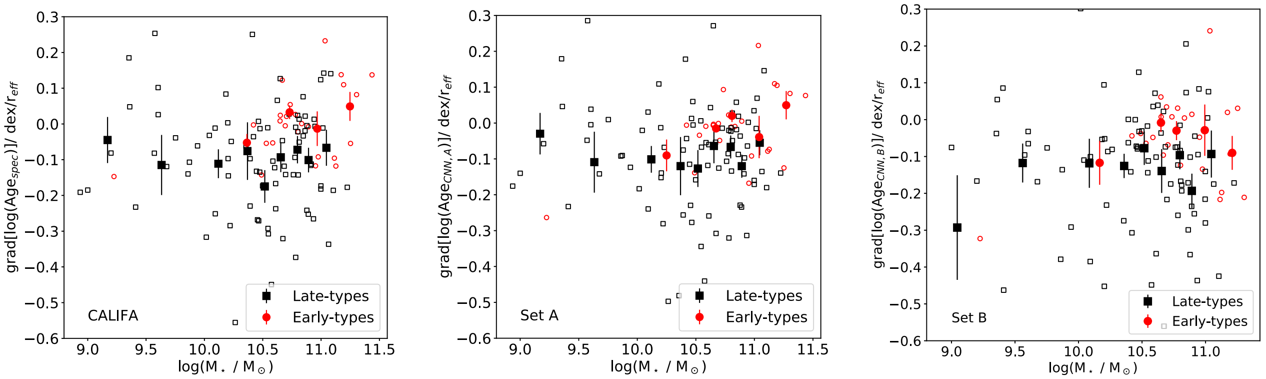


Figure 14. The radial age gradient for a galaxy against its stellar mass, using spectroscopically determined gradients from CALIFA (left), and the gradients calculated from CNN predictions with Set A (middle) and Set B (right). The open red circles (the open black squares) show the values for individual early- (late-) type galaxies. The red-filled circles (the black-filled squares) show the mean value for each bin of 6 (10) galaxies for early- (late-) type galaxies, with error bars showing the standard deviation. This demonstrates the gradients of Set A are more similar to the spectral gradients than those of Set B.

used, in combination with transfer learning (Zhuang et al. 2019), to improve the accuracy of predictions in the future.

ACKNOWLEDGEMENTS

We thank an anonymous referee for their thorough review and helpful detailed suggestions that have improved the manuscript significantly. CLLC, DK, and MS acknowledge the support of the UK's Science and Technology Facilities Council (STFC Grant ST/S000216/1). CLLC is also grateful for the STFC Doctoral Training Partnerships Grant (ST/S50578X/1). IF acknowledges support from the Spanish Ministry of Science, Innovation and Universities (MCIU), through Grant PID2019-104788GB-I00. PSB acknowledges financial support from the Grant PID2019-107427GB-C31 from the Spanish Ministry of Science and Innovation.

DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

Abadi M. et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/
Acquaviva V., 2016, MNRAS, 456, 1618
Baron D., 2019, preprint (arXiv:1904.07248)
Belfiore F., Vincenzo F., Maiolino R., Matteucci F., 2019, MNRAS, 487, 456
Benitez N. et al., 2014, preprint (arXiv:1403.5237)
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Bundy K. et al., 2015, ApJ, 798, 7
Cappellari M., Emsellem E., 2004, PASP, 116, 138
Cenarro A. J. et al., 2019, A&A, 622, A176
Ciucă I., Kawata D., Miglio A., Davies G. R., Grand R. J. J., 2020, preprint (arXiv:2003.03316)
Conroy C., 2013, ARA&A, 51, 393
Croom S. M. et al., 2012, MNRAS, 421, 872
Díaz-García L. A. et al., 2015, A&A, 582, A14
Fabbro S., Venn K. A., O'Briain T., Bialek S., Kielty C. L., Jahandar F., Monty S., 2018, MNRAS, 475, 2978
Folkes S. R., Lahav O., Maddox S. J., 1996, MNRAS, 283, 651
Girardi L., Bressan A., Bertelli G., Chiosi C., 2000, A&AS, 141, 371
González Delgado R. M. et al., 2015, A&A, 581, A103
Hüllermeier E., Waegeman W., 2019, preprint (arXiv:1910.09457)
Kroupa P., 2001, MNRAS, 322, 231
Lovell C. C., Acquaviva V., Thomas P. A., Iyer K. G., Gawiser E., Wilkins S. M., 2019, MNRAS, 490, 5503
Mejía-Narváez A. et al., 2017, MNRAS, 471, 4722
Mendes de Oliveira C. et al., 2019, MNRAS, 489, 241
Ocvirk P., Pichon C., Lançon A., Thiébaut E., 2006, MNRAS, 365, 46
Panter B., Heavens A. F., Jimenez R., 2003, MNRAS, 343, 1145
Pérez-González P. G. et al., 2013, ApJ, 762, 46
Sánchez-Blázquez P., 2016, Stellar Populations of Bulges at Low Redshift. p. 127
Sánchez-Blázquez P., Gorgas J., Cardiel N., González J. J., 2006, A&A, 457, 809
Sánchez-Blázquez P. et al., 2014, A&A, 570, A6
Sánchez S. F. et al., 2012, A&A, 538, A8
San Roman I. et al., 2018, A&A, 609, A20
San Roman I. et al., 2019, A&A, 622, A181
Sarzi M. et al., 2006, MNRAS, 366, 1151
Trager S. C., Faber S. M., Worthey G., González J. J., 2000, AJ, 119, 1645
Vazdekis A., Sánchez-Blázquez P., Falcón-Barroso J., Cenarro A. J., Beasley M. A., Cardiel N., Gorgas J., Peletier R. F., 2010, MNRAS, 404, 1639
Walcher C. J. et al., 2014, A&A, 569, A1
Wenger M. et al., 2000, A&AS, 143, 9
Wisnioski E. et al., 2015, ApJ, 799, 209
Wolf C., Meisenheimer K., Röser H.-J., 2001, A&A, 365, 660
Worthey G., 1994, ApJS, 95, 107
Wu J. F., Boada S., 2019, MNRAS, 484, 4683
Zhuang F., Qi Z., Duan K., Xi D., Zhu Y., Zhu H., Xiong H., He Q., 2019, preprint (arXiv:1911.02685)

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