J-PAS: Measuring emission lines with artificial neural networks

G. Martínez-Solačhe1, R. M. González Delgado4, R. García-Benito1, A. de Amorim2, E. Pérez3, J. E. Rodríguez-Martín1, L. A. Díaz-García14,1, R. Cid Fernandes2, C. López-Sanjuan3, S. Bonoli13,4,5, A. J. Cenarro3, R. A. Dupke6,11,12, A. Marín-Franch3, J. Varela1, H. Vázquez Ramió3, L. R. Abramo7, D. Cristóbal-Hornillos3, M. Moles3,1, J. Alcaniz6,8, P.O. Baqui15, N. Benitez1, S. Carneiro9, A. Cortesi16, A. Ederoclite10, V. Marra15, C. Mendes de Oliveira10, L. Sodré Jr.10, J. M. Vilchez1, K. Taylor11, and JPAS collaboration

1 Instituto de Astrofísica de Andalucía (CSIC), PO Box 3004, 18080 Granada, Spain (e-mail: gimars0@iaa.es)
2 Departamento de Física, Universidade Federal de Santa Catarina, P.O. Box 476, 88040-900, Florianópolis, SC, Brazil
3 Centro de Estudios de Física del Cosmos de Aragón (CEFCA), Unidad Asociada al CSIC, Plaza San Juan, 1, E-44001, Teruel, Spain
4 Donostia International Physics Center (DIPC), Manuel Lardizabal Ibilbidea, 4, San Sebastián, Spain
5 Ikerbasque, Basque Foundation for Science, E-48013 Bilbao, Spain
6 Observatório Nacional, Ministério da Ciência, Tecnologia, Inovação e Comunicações, Rua General José Cristino, 77, São Cristóvão, 29201-460, Rio de Janeiro, Brazil
7 Instituto de Física, Universidade de São Paulo, Rua do Matão 1371, CEP 05508-090, São Paulo, Brazil
8 Departamento de Física, Universidade Federal do Rio Grande do Norte, 59072-970, Natal, RN, Brazil
9 Instituto de Física, Universidade Federal do Bahia, 40210-340, Salvador, BA, Brazil
10 Departamento de Astronomía, Instituto de Astrofísica, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, São Paulo, Brazil
11 Department of Astronomy, University of Michigan, 311 West Hall, 1085 South University Ave., Ann Arbor, USA
12 Department of Physics and Astronomy, University of Alabama, Box 870324, Tuscaloosa, AL, USA
13 Instruments4
14 Academia Sinica Institute of Astronomy & Astrophysics (ASIAA), 11F of Astronomy-Mathematics Building, AS/NTU, No. 1, Sect. 4, Roosevelt Road, Taipei 10617, Taiwan
15 Núcleo de Astrofísica e Cosmologia (Cosmo-ufes) & Departamento de Física, Universidade Federal do Espírito Santo, 29075-910, Vitória, ES, Brazil
16 Observatório do Valongo, Universidade Federal do Rio de Janeiro, 20080-090, Rio de Janeiro, RJ, Brazil

Received ? / Accepted ?

ABSTRACT

The detection and measurement of emission lines is important to understand the evolution of galaxies through cosmic time. Throughout this paper we present a new method to carry out this task. J-PAS will observe 8000 deg² of the northern sky in the upcoming years with 56 photometric bands. The release of such amount of data brings us the opportunity to employ machine learning methods in order to overcome the difficulties associated with photometric data. We aim to detect and measure emission lines in J-PAS up to \( z \approx 0.9 \). We used Artificial Neural Networks (ANNs) trained and tested with synthetic J-PAS photometry from CALIFA, MaNGA, and SDSS spectra. We carry out two tasks: firstly, we cluster galaxies in two groups according to the values of the equivalent width (EW) of \( \text{H}_\alpha \), \( \text{H}_\beta \), \([\text{NII}]\)6584, and \([\text{OIII}]\)5007 lines measured in the spectra. Then, we train an ANN to assign to each galaxy a group. We are able to classify them with the uncertainties typical of the photometric redshift measurable in J-PAS. Secondly, we utilize another ANN to determine the values of those EWs. Subsequently, we obtain the \([\text{NII}] / \text{H}_\alpha\), \([\text{OIII}] / \text{H}_\beta\), and O 3N 2 ratios recovering the BPT diagram \((\text{OIII}) / \text{H}_\beta\) vs \([\text{NII}] / \text{H}_\alpha\). We study the performance of the ANN in two training samples: one is only composed of synthetic J-PAS photo-spectra (J-spectra) from MaNGA and CALIFA (CALMa set) and the other one is composed of SDSS galaxies. We can reproduce properly the main sequence of star forming galaxies from the determination of the EWs. With the CALMa training set we reach a precision of 0.101 and 0.091 dex for the \([\text{NII}] / \text{H}_\alpha\) and \([\text{OIII}] / \text{H}_\beta\) ratios in the SDSS testing sample. Nevertheless, we find an underestimation of those ratios at high values in galaxies hosting an active galactic nuclei. We also show the importance of the dataset used for both training and testing the model. ANNs are extremely useful to overcome the limitations previously expected concerning the detection and measurements of the emission lines in surveys like J-PAS. Finally, we compare the properties of emission lines in galaxies observed with miniJPS and SDSS. Despite of the limitation of such a comparison, we find a remarkable correlation in their EWs and show the capability of the method to detect galaxies with EWs greater than 3 Å.

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

1. Introduction

The study of the formation and evolution of galaxies through cosmic time has been addressed in the last decades by understanding how their physical properties leave footprints in the spectral energy distribution (see e.g. Díaz-García et al. 2019, and references therein). Both the analysis of the light coming from stars and the ionized interstellar gas can be converted by well-known recipes to physical quantities such as the stellar mass, star formation rate (SFR), dust attenuation,
luminosity-age, gas-phase metallicity or can unveil the main ionization mechanism responsible for the optical emission lines we observe in the spectrum (for some of the most recent reviews on these topics, see Conroy 2013; Madau & Dickinson 2014; Kewley et al. 2019).

The most massive and youngest stars within galaxies are responsible for the ultraviolet emission in the spectrum, but many times the presence of dust grains does not allow ultraviolet photons to travel freely through the interstellar medium and consequently makes it difficult to constrain the SFR from the blue part of the spectrum alone. However, those stars can actually ionize the surrounding interstellar gas. Very rapidly, hydrogen atoms recombine leaving tracks in form of emission lines at a particular wavelength in the spectrum. The Balmer series places Hα at 6562.8 Å, hence it is less affected by dust extinction and an excellent tracer to measure SFRs up to z ~ 0.4 in the optical range (Catalán-Torrecilla et al. 2015). Other lines, such as the forbidden [OIII]λλ4959, 5007 Å and [NII]λλ6548, 6584 Å doublets, are sensitive to the gas-phase metallicity, which is ideal for investigating the metal enrichment of gas throughout cosmic time (Maiolino & Mannucci 2019). The [NII]λ6584/Hα and [OIII]λ5007/Hβ ratios among others are used to construct the so-called BPT diagrams (Baldwin et al. 1981), which distinguish galaxies where the gas has been ionized due to the presence of an active galactic nuclei (AGN) from those where the main ionization mechanism comes from high rates of star formation in the galaxy or shock ionized gas regions.

Even though spectroscopic surveys revolutionized astronomy in many fields, they provide a limited picture of the universe in many senses. Both Multi-Object Spectroscopy and integral fields units (IFUs) surveys are partially biased due to pre-selected samples where some properties such as fluxes, redshift or galaxy-size are limited to a certain range. Some of these issues can partially be solved with narrow band photometric surveys. Although they have been historically limited to few filters, they can act as low-resolution spectrographs and they are able to map the sky quickly and deeply; therefore, giving a more comprehensive snapshot of the universe. Needless to say, some astrophysical analyses will always require the highest possible spectral resolution to fully exploit all the information encoded in the spectrum.

Maybe one of the most competitive astrophysical surveys designed to overcome the weakness of photometry and spectroscopy, halfway between them, is the Javalambre-Physics of the Accelerating Universe (J-PAS, Benitez et al. 2014). It will sample the optical spectrum with 56 narrow-band filters for hundreds of millions of galaxies and stars over ~ 8000 deg². This is equivalent to a resolving power of R ~ 50 (J-spectrum). Initially thought to explore the origin and nature of the dark energy in the universe, J-PAS is also useful for galaxy evolution studies and to detect emission line objects (Bonoli et al. 2020). However, the large number of galaxies peaking over a wide range of redshift makes it difficult to employ traditional methods such as subtracting from the emission line flux the image of the stellar continuum (Vilella-Rojo et al. 2015). Furthermore, line fluxes will contribute to several J-PAS filters which also vary with the redshift of the object. Consequently, it is necessary to develop new techniques and algorithms in order to leverage completely the capability of J-PAS.

Machine learning techniques have effectively become a powerful tool over many fields where large quantities of data are available. The capability of these algorithms to find patterns in the data without making any empirical or theoretical assumption turns out to be their main advantage. In the last decades, astrophysical surveys are increasingly releasing vast amounts of data, which brings the opportunity of employing the most sophisticated up-to-date algorithms in order to analyse them faster and more efficiently. The applications range from the estimation of photometric redshifts (Pasquet et al. 2019; Cavuoti et al. 2017), identification of stars (Whitten et al. 2019), classification of galaxies (Dominguez Sánchez et al. 2018), separation between galaxies and stars (Baqui et al. 2020) to the determination of the SFR (Delli Veneri et al. 2019; Bonjean et al. 2019) to cite some of the most recent research. In this work, we developed a new method based on Artificial Neural Networks (ANN) to detect and measure some of the main emission lines in the optical range of the spectrum: Hα, Hβ, [NII]λ6584, and [OIII]λ5007.

This paper is organized as follows. We present in Sect. 2 J-PAS data together with data from other surveys that have been used to train and test the ANNs. In Sect. 3 we describe in detail the main characteristics of the ANNs, how they can be trained and tested to deal with the uncertainties associated to the data. In Sect. 4 we show the performance of ANNs in SDSS simulated data sets and discuss its main weakness. In Sect. 5 we test our method in galaxies observed both in miniJ-PAS and SDSS. Finally, we summarize in Sect. 6 and point out the steps needed to improve and extend the performance of the ANN in detecting and measuring emission lines.

2. J-PAS and spectroscopic data

In this section we present J-PAS and the spectroscopic data used throughout this paper for training and testing the model.

2.1. J-PAS

J-PAS is an astrophysical survey (Benitez et al. 2014) planning to map ~ 8000 deg² of the northern sky with 56 bands. This is, 54 narrow-band filters in the optical range plus 2 mediumband, one in the near-UV and another in the NIR. With a separation of 100 Å, each narrow-band filter will have a FWHM of ~ 145 Å. The observations will be carried out with the 2.55 m telescope (T250) at the Observatorio Astrofísico de Javalambre, a facility developed and operated by CEFCA, in Teruel (Spain) using the JPACam, a wide-field 14 CCD-mosaic camera with a pixel scale of 0.2267 arcsec and an effective field of view of ~ 4.7 deg² (see Cenarro et al. 2019; Taylor et al. 2014; Marin-Franch et al. 2015). The survey is expected to detect objects with an apparent magnitude equivalent to 14.6 mag (Bonoli et al. 2015). The survey is expected to detect objects with an apparent magnitude equivalent to 14.6 mag (Bonoli et al. 2015). The survey is expected to detect objects with an apparent magnitude equivalent to 14.6 mag (Bonoli et al. 2015). The survey is expected to detect objects with an apparent magnitude equivalent to 14.6 mag (Bonoli et al. 2015). The survey is expected to detect objects with an apparent magnitude equivalent to 14.6 mag (Bonoli et al. 2015).

The J-PAS project started its observations taking data with the Pathfinder camera observing four AEGIS fields with 60 optical bands amounting to 1deg². These data allow us to build a complete sample of galaxies up to $z_{SDSS} < 22.5$ mag (Bonoli et al. 2020). More than 60,000 objects have been detected and can be downloaded from the website of the survey¹. A detail description of the survey, referred as to miniJ-PAS, is developed in Sect. 5.1

¹ http://www.j-spas.org/
acubes Sánchez et al. (2016a,b). The analysis of the stellar populations and ionized gas provides spatially-resolved information of the strongest emission lines in the optical range for a total of 4670507 spaxels from 2755 galaxies.

2.4. SDSS survey

The Sloan Digital Sky Survey (SDSS, York et al. 2000) contains spectroscopic measurements for more than three million astronomical objects and deep images of one third of the sky in five optical bands. The spectra were taken with a fiber of 3” in diameter and a spectral coverage of 3800–9200 Å at a resolution of $R \sim 2000$. We use here the publicly available MPA-JHU DR8 catalog, from the Max Planck Institute for Astrophysics and the Johns Hopkins University (Kauffmann et al. 2003b; Brinchmann et al. 2004). All the information regarding the catalog and the fitting procedure of the galaxy physical properties can be consulted online 3. The catalog provides a total of 818333 galaxies with redshift up to $z \sim 0.35$. We take only galaxies with reliable emission line measurements. As described in the data-model of the catalog, we can do that by excluding from the sample objects with RELIABLE = 0 and/or WARNING > 0. We also discard galaxies where J-PAS synthetic magnitudes cannot be calculated due to the lack of data in certain wavelength range of SDSS spectra. Finally, we end up with 701975 galaxies.

3. Method of analysis.

In this section we describe the architecture of the network in Sect. 3.1 and the strategies used for training and testing the model in Sect. 3.2. We also explain how to deal with photo-redshift uncertainty in Sect. 3.3, how errors can be estimated in Sect. 3.4, and how to treat missing data in Sect. 3.5.

3.1. Architecture of the Network

In this paper we use a class of ANN called fully connected neural network. The implementation has been made with TensorFlow (Abadi et al. 2015) and Keras libraries (Chollet et al. 2015) in Python. It is composed of a set of layers which have a specific number of neurons. The first layer contains the inputs of the network. In our application, the inputs are the colors of J-PAS measured with respect to the filter corresponding to $H\alpha$ for each spectrum. For instance, in nearby galaxies ($z < 0.015$) $H\alpha$ emission line will be captured by the J0660 band. Then, the color in the filter $Ji$ is defined as the difference respect to the magnitude measured in the J0660 band ($C_i = m_{AB}(\lambda0660) - m_{AB}(Ji)$). The final layer contains the output of the network, sometimes also named targets in the machine learning argot. Our targets are the equivalent width (EW) of $H\alpha$, $H\beta$, $[NII]$,6584 and $[OIII]$,4959,5007. We built two different ANNs: one performs a regression task and obtains the values of these EWs, this network will be referred to as ANN$_R$. The other, ANN$_C$, carries out a classification between galaxies without emission lines (below a given threshold) and emission line galaxies by imposing cuts in the EWs of the mentioned lines. We could have performed this classification based on the values yielded by the ANN$_R$ but an algorithm specifically constructed for that will always obtain better results.

As we mentioned before, emission line fluxes have contribution to different bands according to the redshift of the source

---

2 https://www.sdss.org/dr14/manga/manga-data/manga-pipe3d-value-added-catalog/

3 www.sdss3.org/dr10/spectro/galaxy_mpajhu.php
and the width of the emission line. The redshift might be treated as an input in the model but that would imply to train the ANN with a uniform distribution in this parameter, otherwise the ANN would not be able to predict equally at all redshifts. Furthermore, this approach would reduce our sample size and limit our range of predictability due to the different redshift coverage of CALIFA, MaNGA and SDSS. For these reasons, we train a different ANN for each redshift, going from 0 to 0.35 with a step of 0.001. We shift all the spectra of the training set in wavelength at the same redshift and we compute the colors within the common wavelength range between J-PAS and the spectroscopic surveys described in Sect. 2. This range depends on the redshift and consequently the number of inputs vary between 28 and 39 colors.

Between the input and the output layers the ANN can hold inner layers, commonly called hidden layers, with absolute freedom to decide the number of layers and neurons in it. There is no standard recipe to find the optimal architecture of a network, but one hidden layer is sufficient for the large majority of problems and usually the optimal amount of neurons in the hidden layer varies between the size of the input and the size of the output layers. Our ANNs have 20 neurons in the hidden layer, which is in between the number of inputs (34 colors in average) and the number of outputs (four EWs for the ANN R and two classes in the case of the ANN C). A schematic view of the ANN R used in this work can be seen in Fig. 2.

All the neurons in a given layer are connected to the neurons in the contiguous layer by a matrix of weights W and a bias B:

\[ L_n = g(W_n \cdot L_{n-1} + B_n) \]  

(1)

where \( L_n \) refers to layer n. \( L_0 \) and \( B_0 \) are the inputs of the ANN and g is the activation function of neurons. It worth mentioning the importance of such function, being responsible for the non-linear behavior in the network. Otherwise, the outputs would be simply a linear combination of the inputs, which would not be sufficient to address most of the problems. We use the so-called Rectified Linear Unit (ReLU) activation function (Nair & Hinton 2010). Typically, ANN are trained using a supervised learning algorithm commonly referred to as backpropagation. Adjusting the set of weights and bias that minimizes a certain loss-function is the actual process of training. Vanishing gradients do not let the algorithm to know in which direction the parameters should be readjusted to find the global minimum of the loss-function. For this reason, the derivative of the activation function of neurons should not go to zero easily. This is the case of ReLU, which at least in half of the times will not vanish4. For regression-like problems the most common loss-function is usually a mean square error, while for binomial classification the binary cross entropy is frequently employed. We make use of these functions in our models.

One important aspect to take heed of when we are training an ANN is to avoid overfitting. Improving the loss-function indefinitely will make the algorithm to fit features of the data that do not represent the general trend. Consequently, the predictability of the network will be compromised. We can avoid that by imposing a maximum value over the weights that each neuron can carry.

\[ \frac{d g(x)}{dx} = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases} \]

Article number, page 4 of 19

Optimising the architecture of the network is a process that requires tweaking many parameters and there is always room for improvement. Most probably, other configurations could obtain similar results or even marginally ameliorate the performance of the network. Nevertheless, this escapes the scope of the present paper. We rather focus on satisfying our scientific goals detailed in Sect. 4.

3.2. Training strategy

We generate synthetic J-PAS data by convolving the spectra presented in Sect. 2 with the J-PAS filter system. Since CALIFA, MaNGA, SSDS, and J-PAS have different wavelength coverage, we only use in our model the common wavelength range of the four instruments at \( z = 0 \), which is 3810 – 6850 Å.

The training sample is built differently depending on whether we are dealing with a classification or a regression task. Selecting the most appropriate one is the cornerstone of any machine learning problem. The training sample should be representative of the target one, that is, as close as possible to J-PAS data, and complete enough to make successful predictions. In this regard, many aspects need to be taken into account. Firstly, in order to ensure the algorithm receives the most reliable information, one would desire to select only the spectra where emission lines have been measured with high signal-to-noise ratio. However, being too strict in the selection criterium induces a bias towards line-emitting galaxies and reduces significantly the size of the sample. Secondly, while CALIFA and MaNGA have observed the nearby universe resolving spatially the physical properties of the interstellar medium within galaxies, SDSS can only see the inner parts of nearby galaxies but with the advantage of covering distances further away in the universe. It has been shown how spatial resolution affects the location of points (spaxels) in the BPT, possibly altering AGN classification and/or simulating it via mixed spectral featured (Gomes et al. 2016). Furthermore, the emission lines catalogs obtained from these surveys have been derived with different fitting tools which makes it difficult to compare them in equal terms. In essence, there is not a simple and unique way of putting together all these data and build the training set that better represents the universe as J-PAS will look at it. Instead, we propose to train the ANN with different training sets in order to understand the source of errors and inaccuracies of the model.

![Schematic diagram of the ANN used for predicting lines emission at rest frame.](Image)

Fig. 2. Schematic diagram of the ANN used for predicting lines emission at rest frame. The J0660 filter is our reference band for colors.
3.2.1. Training and testing sets in the ANN

With the aim of identifying galaxies with low and high emission lines, we train the ANN to perform a binary classification based on the EW of \( H\alpha, H\beta, [NII],6584 \) or \([OIII],5007\). Galaxies in the training set are classified according to the following criteria:

\[
EW(H\alpha) > EW_{\text{min}} \quad \text{or} \quad EW(H\beta) > EW_{\text{min}} \quad \text{or} \quad EW([OIII]) > EW_{\text{min}} \quad \text{or} \quad EW([NII]) > EW_{\text{min}} \Rightarrow \text{Class 1}
\]

where \( EW_{\text{min}} \) is the EW used as threshold and takes the following values: 3, 5, 8, 11, and 14 Å. In short, if a galaxy has an EW greater than the threshold in any of these lines, it will be considered as Class 1. If all the EWs in a galaxy are below the threshold then it will be tagged as Class 2. In most of the cases \( H\alpha \) is the most powerful emission line and consequently it will decide whether galaxies belong to one class or other. This type of classification allows us to disentangle the structure of the bi-modal distribution found in the EW of \( H\alpha \) in CALIFA and SDSS galaxies (Bamford et al. 2008; Lacerda et al. 2018) and will help us to better discern between galaxies with and without emission lines in J-PAS. Since such classification seems to be an easy task for the ANN\(_C\), the combination of data from different surveys explored in this work does not improve or worsen its performance. Consequently, for the sake of simplicity, we train only with CALIFA synthetic J-spectra and we test with SDSS galaxies. We do not impose any cut in the errors of the EWs but we ensure to have the same amount of J-spectra in both classes in the training set. We end up with 200000 synthetic J-spectra to perform the training.

3.2.2. Training and testing sets in the ANN\(_R\)

For the purpose of obtaining the values of the EWs of galaxies in J-PAS, we propose two training sets. The first one, what we call the CALMa set, is only composed of CALIFA and MaNGA synthetic J-spectra while the second one, the SDSS set, includes only SDSS galaxies. We test the performance of the model by removing randomly 15000 synthetic J-spectra from the training samples: 5000 from CALIFA, 5000 from MaNGA and 5000 from SDSS. Those synthetic J-spectra are considered as validation or test samples depending on the training sample. For instance, if we train with the CALMa set, we use MaNGA and CALIFA samples to tune the parameters of the model (validation samples) and SDSS galaxies to actually evaluate the model; and the other way around: if we train with the SDSS sample, SDSS galaxies plays the role of the validation sample and CALIFA and MaNGA synthetic J-spectra are used for testing purpose. In this way, we ensure that the model does not get biased due to the systematics associated to the fitting tools and/or the ones associated to the instruments.

We add to the training set only those synthetic J-spectra where emission lines have an error below a certain threshold. In the case of MaNGA galaxies, spaxels with signal-to-noise-ratio (S/N) below 10 in the flux of \( H\alpha, H\beta, [NII],6584 \) or \([OIII],5007\) are discarded. However, we were more flexible with Voronoi zones in CALIFA and SDSS galaxies, going down to a S/N of 2.5. Such flexibility allows us to increase the amount of low-emitting galaxies in the samples. In addition, when it comes to the CALMa set, we achieve a more equilibrated weight between the prominence of CALIFA and MaNGA in the training sample. We also exclude from the training set the galaxies where the EWs are greater than 600 Å. Since the loss function is proportional to the EWs, those galaxies force the ANN\(_C\) to fit at the same time two antagonistic regimes: low-emitting and extreme emission line galaxies. Consequently, it would worsen the performance of the ANN\(_R\) in the range of interest. Finally, we end up with a training set of 134000 synthetic J-spectra from CALIFA, 280270 from MaNGA, which together form the CALMa set; and 135300 galaxies in SDSS set.

3.3. Photo-redshift uncertainty

Even though J-PAS will provide redshifts with high precision (Benitez et al. 2014, \( \Delta z \leq 0.3\% \) for luminous red galaxies), the performance of the ANN could be compromised in many cases. Let us assume for example that we aim to compute the EWs of a galaxy at redshift 0.3 with \( \Delta z = 0.003 \). In the best case scenario, the galaxy redshift would be between 0.296 and 0.304. According to our redshift bin, we have 8 possible ANNs to try with. While in the vicinity of the true redshift the ANN can reasonably make a good job, in the extremes the EWs would dramatically be underestimated. Since colors are computed with respect to a filter far away from the one corresponding to \( H\alpha \), the ANN will interpret as an absorption line what indeed is an emission line. Although the probability density functions (PDFs) of the photo-z can help improving the predictability assigning weights to each redshift, whenever we found a non-gaussian PDF with, for instance, an asymmetric distributions with two peaks, it would be difficult for the ANN to make reasonable predictions. One solution is to consider only the configurations (redshifts) that maximizes a certain function. Certainly, for emission line galaxies, the redshift where the sum of all EWs reaches the highest value is close to the true redshift. However, this redshift overestimates the EWs in galaxies with low emission. In order to minimize such effect, we average over the five configurations (redshifts) that maximize the sum of all EWs within the photo-redshift uncertainty. As we discuss later in Sec. 4.4, this method is able to somehow recompute the distance of the galaxy correcting a possible deviation from the spectroscopic redshift in galaxies where \( \sum EW_i > 20 \text{ Å} \). Therefore the method of the five maximum, hereafter 5max, can certainly help the ANNs to improve its performance but cannot be used with the ANN\(_C\). Most probably, it would increase the amount of false positives as the redshift uncertainty increases. In Sect. 4 we quantify how the error in the redshift can impact the predictions of the ANN\(_C\) and the ANN\(_R\). Fortunately, the ANN\(_C\) is less sensitive to that (see Fig. 3 and Table 1).

3.4. Estimation of errors

The uncertainty of the ANN method can be estimated considering three sources of errors: the error of the photometry, the error in the photometric redshift, and the intrinsic error of the ANN training. Before the training actually starts, weights and biases in ANN can be set to a certain value by initialising randomly according to any distribution function. Generally, each initialization state will converge to different local minimum of the loss-function. Even though it is possible to find the state

\[ \Delta z = z - z_{\text{photo}}. \]
that leads to the best score over the validation sample, usually a Monte Carlo approach called the committee, this is, the mean of the individual predictions of a set of ANN, will be a more robust and accurate estimate of the targets. Then, the variations of the outputs in each individual member of the committee respect to the mean provide an estimation of the uncertainty in the predictions intrinsically associated to the training procedure. The list bellow details the steps to follow in order to account for the contribution of each uncertainty to the errors budget.

1. Photometric error: we input the ANN with \( N + 1 \) different values of the magnitude, where one corresponds to the nominal value and the other \( N \) are randomly drawn from a gaussian distribution centred on the nominal value and with standard deviation equal to the photometric error. The median (M) and the median absolute deviation (MAD) of \( N+1 \) predictions give us the prediction and the weight of one member in one committee:

\[
P_{i|z} = M[P_{0i}, P_{1i}, ..., P_{N+1i}]
\]

\[
W_{i|z} = 1/MAD[P_{0i}, P_{1i}, ..., P_{N+1i}]
\]

where \( i \) stands for the committee member and \( z \) for the redshift.

2. ANN intrinsic error: the prediction of the committee in a given redshift can be estimated by computing the average (AVG) of all members in the committee with the weights obtained above. The error of the committee is simply the MAD respect to the average:

\[
\epsilon_{i|z} = M[|P_{zi} - P_{0zi}|, |P_{zi} - P_{1zi}|, ..., |P_{zi} - P_{Nzi}|]
\]

where \( m \) refers to the number of members in the committee. We found that averaging over five members is enough to obtain reliable results.

3. Photo-redshift uncertainty: we compute the median value of \( n \) committees, one for each redshift. In the case of the ANN we select the five maximum setting (see Sect. 3.3) and for the ANNc we consider all the redshift within the error range.

\[
P_{ANN} = M[P_{x0}, (max_x), P_{x1}, (max_x), ..., P_{x4}, (max_x)]
\]

Finally, the error is the quadratic sum of the median error of all committees plus the dispersion of these committees respect to the median, which gives us the contribution of the redshift uncertainties.

\[
\epsilon_{ANN} = \sqrt{M[\epsilon_{i|z}^2, \epsilon_{i|z}^2, ..., \epsilon_{i|z}^2]} + MAD[P_{x0}, P_{x1}, ..., P_{x4}]
\]

If the spectroscopic redshift of the object were known, the expression above would be simply:

\[
\epsilon_{ANN} = \epsilon_{spec}
\]

3.5. Missing data

Many are the problems, both related to the data reduction or the observation, that could lead to incomplete or missing data in some of the filters. Consequently, a small fraction of our sample will lack photometric measurements in some of the filters used by the ANN. Certainly, many of such objects will have to be rejected automatically if the photometry is not reliable in the bands capturing the emission lines. However, in some of them, the photometry might be problematic only in some of the filters dominated by the stellar continuum. One solution requires training several ANN considering different configurations where part of the data is missing. Nevertheless, this would imply testing the performance of the ANN in many scenarios and would be computationally very expensive. The other solution is to replace the missing data in the corresponding filter with the fluxes obtained from the spectral fitting of the stellar continuum. Several spectral fitting codes can be used such as MUFFIT (Díaz-García et al. 2015) or BaySeAGal (Amorim et al. in prep.). This analysis provides reliable photometric predictions for the missing data, as well as information regarding their stellar population properties (e.g., stellar mass, age, and extinction, which is always necessary for a more comprehensive picture). Furthermore, the stellar continuum is needed for obtaining absolute emission line fluxes. We follow this technique to treat the missing data in J-PAS.

4. Validation of the method.

In this section we perform several tests to study the predictability and limitations of the model. Firstly, we evaluate the capability of the ANNc in Sect. 4.1. Secondly, in Sect. 4.2, we compare the predictions of the EWs obtained by the ANN and trained with the CALMa set with the SDSS testing sample. Then, in Sect. 4.3 we compare the performance of the different training sets proposed in Sect. 3.2.2. Finally, we study in Sect. 4.4 the impact of the redshift uncertainty on the ANN as a function of the EW.

4.1. Classifying galaxies

The ANNc is trained with the CALIFA training sample. For evaluating its efficiency, we randomly select two samples from the SDSS catalog: 5000 galaxies belonging to Class1 and 5000 to Class2. (see Sect. 3.2.1). For each galaxy the ANN yields a number between 0 and 1 indicating the probability of being one of the two classes. As we discuss in Sect. 4.4, the \( 5_{max} \) method (Sect. 3.3) is not suitable for galaxies without emission lines. Most probably, it would increase the amount of false positives as the redshift uncertainty increases. Since we have noticed that the ANNc is less sensitive to redshift and is able to classify galaxies even when its uncertainty is high, we simply compute the average of each one of the predictions within the redshift interval defined by \( \delta z \).

We show in Fig. 3 the receiver operating characteristic (ROC) curve, which represents the true positive rate (TPR) versus the false positive rate (FPR) for \( EW_{min} = 3 \) Å. We also show how the ROC curve varies as a function of the redshift uncertainty. The ANNc scores very high even when \( \delta z = 0.01 \) and loses efficiency gradually as the uncertainty in the redshift increases. We summarize in Table 1 the area under the ROC curves for others threshold settings. The ROC curves do not show remarkable changes in function of the threshold used in the classification.

4.2. Emission-line galaxies: EWs, line ratios and BPT diagram

In this section we discuss how the CALMa training set (see Sect. 3.2.2) scores in the SDSS testing sample. We use the
Blue dashed line shows the performance of a random classifier.

Table 1. Area under the ROC curve as a function of the redshift uncertainty and the threshold used in the classification.

| EW  | Area (Δz = 0.01) | Area (Δz = 0.02) | Area (Δz = 0.03) |
|-----|------------------|------------------|------------------|
| 1 A | 0.9949           | 0.9629           | 0.8920           |
| 3 A | 0.9948           | 0.9507           | 0.8920           |
| 8 A | 0.9938           | 0.9604           | 0.819            |
| 11 A| 0.9915           | 0.9594           | 0.8920           |
| 14 A| 0.9894           | 0.9600           | 0.8921           |

spectroscopic redshift provided in the catalog without considering any error so as to separate the uncertainties intrinsically associated to the model from those related to redshift. We do not consider the errors of SDSS spectra, we rather add gaussian noise to each magnitude 100 times assuming an average S/N of 10. This allows us to treat all galaxies in the same manner and assume higher errors.

The testing set from CALIFA, MaNGA, and SDSS are composed of 5000 synthetic J-spectra with S/N in the EWs above 10. This criterion excludes many galaxies with low-ionization nuclear emission-line region (LINER). We also exclude galaxies where the EWs are greater than 600 Å to test the model in the range of which we trained the \( \text{ANN}_R \). We note that one of the greatest limitations of any machine learning algorithm is the incapability to make trustworthy predictions outside the training parameter space. Hence, even though we are able to identify strong and weak emission lines galaxies, their EWs might not be accurate due to these selection criteria on the training sample.

4.2.1. Equivalent widths

Fig. 4 compares the EWs predicted by the \( \text{ANN}_R \) and those in the SDSS testing sample (extracted from the MPA-JHU DR8 catalog). We do not plot the errors yielded by the \( \text{ANN}_R \) for visual reasons. We obtained on average an error of 3% for \( \text{H}_\alpha \) and \( \text{H}_\beta \), 5% for [NII]λ6584, and 10% for [OIII]λ5007. The plots on the left are color-coded with the density of points and the ones in the middle with the redshift of the galaxy. The histograms on the right represents the relative difference between the \( \text{ANN}_R \) predictions and the SDSS testing set. We constrain better the EW of \( \text{H}_\alpha \) followed by \( \text{H}_\beta \), [OIII]λ5007 and [NII]λ6584 (see median and median absolute deviation in Fig. 4). The \( \text{H}_\alpha \) line, which is the most powerful one, presents less dispersion and bias. \( \text{H}_\beta \) and [OIII]λ5007 lines are recovered with similar precision and [NII]λ6584 line show more dispersion and bias. We observe that [NII]λ6584 line saturates at high values, that is to say, the EWs tend to be underestimated as the strength of the line increases. The same effect occurs in the [OIII]λ5007 line in form of a second branch. We analyze this effect in Sect. 4.2.2. We do not observe strong color gradients in the plots color-coded with the redshift, indicating we are not biased regarding the distance of the objects.

In summary, the EWs of \( \text{H}_\alpha \), \( \text{H}_\beta \), [NII]λ6584, and [OIII]λ5007 can be predicted with a relative standard deviation of 8.7%, 14.3%, 15.9%, and 16.4% respectively. \( \text{H}_\alpha \), \( \text{H}_\beta \), [NII]λ6584, and [OIII]λ5007 lines presents a relative bias of 0.17%, 5.4%, 4.8%, and –6.4% respectively. In a future work, we will study the distribution of all these values using a real and complete sample of galaxies from miniJ-PAS.

4.2.2. Ratios between emission lines

From the EWs we can easily obtain the ratios of [NII]λ6584/\( \text{H}_\alpha \) and [OIII]λ5007/\( \text{H}_\beta \) under the approximation that each couple has the same stellar continuum. From that, we also obtain the metallicity indicator \( O_3 N_2 \) (Pettini & Pagel 2004). Fig. 5 shows the comparison between the logarithmic ratios obtained with \( \text{ANN}_R \) and the SDSS testing sample. As in Fig. 4 the plots are color-coded with the density of points (left column) and the redshift of the galaxy (middle panel). The histograms on the right show the logarithmic difference between the \( \text{ANN}_R \) predictions and the SDSS testing set.

The [NII]/\( \text{H}_\alpha \) ratio is predicted within 0.10 dex and a bias of –0.028 dex. The [OIII]/\( \text{H}_\beta \) ratio is slightly better constrained, with no bias and a dispersion of 0.091 dex. Finally, the \( O_3 N_2 \) is recovered within 0.112 dex and a bias of 0.038 dex. The saturation of the [NII]λ6584 line at high values is responsible of the same effect observed in the [NII]/\( \text{H}_\alpha \) ratio. Since MaNGA and CALIFA surveys are mainly composed of star-forming galaxies, the \( \text{ANN}_R \) has few spectra to constrain the ratio of [NII]/\( \text{H}_\alpha \) in galaxies hosting an AGN. To a lesser extent, that also occurs as well in the [OIII]/\( \text{H}_\beta \) ratio for galaxies with values higher than 3.2 and in form of a second branch in the [OIII]λ5007 line.

4.2.3. BPT diagram

In Fig. 6 we compare the BPT diagram recovered by the \( \text{ANN}_R \) (left plot) and the one obtained from the SDSS testing sample (right plot). Galaxies are color-coded with the density of points and are grouped into four classes by three dividing lines: star-forming, composite, Seyfert, and LINER. The solid curve is derived empirically using the SDSS galaxies (Kauffmann et al. 2003a, hereafter ka03). The dashed curve is determined by using both stellar population synthesis models and photoion-
Fig. 4. EWs of Hα, Hβ, [NII]λ6584 and [OIII]λ5007 predicted by the ANN_R compared to SDSS testing sample. The ANN_R is trained with the CALMa set. The color-code represents the probability density function defined by a Gaussian kernel (right panel) and the redshift of the objects (left panel). The histograms are normalized to one and show the relative difference between both values. Black and blue numbers are the median and the median absolute deviation of the difference. Black and grey dashed lines on the left are lines with slope one and the best linear fit respectively. We perform a sigma clipping fit with $\sigma = 3$ to exclude outliers. Red dashed line represents the median.

Article number, page 8 of 19
Fig. 5. Comparison between $[\text{NII}]/H\alpha$, $[\text{OIII}]/H\beta$ and O 3N 2 ratios estimated by the ANNe and SDSS testing sample. Same scheme of Fig. 4. The ANNe is trained with the CALMa set.

In order to investigate the reason why star-forming galaxies populate more the BPT diagram obtained with the ANNe, we show in Fig. 7 the direction towards the location where galaxies should be placed in the BPT according to SDSS MPA-JHU DR8 catalog. The vectors are color-coded with the distance of each galaxy between the two BPT diagrams and more distance ones are plotted last. On average, star-forming galaxies deviate 0.10 dex while Seyfert and composite galaxies do 0.12 dex. On the right panel of Fig. 7, we plot the angular distribution of star-forming, Seyfert and composite galaxies. The angle is defined as a clockwise rotation towards the $x$ axis. While star forming galaxies do not show any preferential
direction, Seyfert and composite galaxies point with an average angle of 45° in the diagram.

### 4.3. Comparison between different \( \text{ANN}_{\alpha} \) training sets

As we pointed out in the Sect. 3.2.2 we have trained the \( \text{ANN}_{\alpha} \) with two different training samples. In the Appendix A we show the results obtained with the SDSS training set in the SDSS testing sample. A quick look at these plots (Appendix A.1, A.2 and A.3) proves the importance of testing the model on data with a different observational setup and calibration. Considering the fact that the EWs are estimated from a pseudo-spectrum (J-spectrum) with a much lower resolving power, the performance of the SDSS training set in SDSS testing sample is outstanding. Nevertheless, it would not be realistic to deduce from that the actual capability of this method to predict in J-PAS data. Testing the CALMa training set with SDSS galaxies or vice versa gives us a better picture of the weakness and inaccuracies of the model. For instance, the predictions made by \( \text{ANN}_{\alpha} \) trained with SDSS set on the \([\text{NII}]/\text{H}\alpha\) and \([\text{OIII}]/\text{H}\beta\) ratios of MaNGA and CALIFA spaxels tend to be overestimated. This is the opposite effect observed when the \( \text{ANN}_{\alpha} \) is trained with CALMa training set and tested on SDSS galaxies. The performance on the validation samples, that is, the data that belongs to the same survey, is generally better. For the sake of illustrating the performance of both training sample (SDSS test and CALMa set) in each one of the testing sets (CALIFA, MaNGA and SDSS) we create a comparison table (Table 2). As it can observe, there will always be a line that is better recovered in one particular simulation, for example \( \text{H}\alpha \) in CALMa vs SDSS, but the overall performance of the \( \text{ANN}_{\alpha} \) is generally more accurate with data from the same survey.

#### 4.4. Dependency on the equivalent width and redshift uncertainty

A simple test to confirm the capability of the \( 5_{\text{max}} \) method to retrieve the redshift of the object is to verify whether the average redshift over the five configuration is far from the true redshift. Normally, we would compute the EWs only in the redshift within the PDF of photo-zs before applying the \( 5_{\text{max}} \), but let us assume we do not have any information regarding the redshift of the object. Then, we have to calculate the EWs in all the redshift from 0 to 0.35 inside the grid and pick only the five redshifts that maximize their sum. Fig. 8 shows this scenario where points are color-coded with the spectroscopic redshift. For emission line galaxies \( \sum \text{EW}_{i} > 20 \, \text{Å} \), this method is able to obtain the redshift of the object with high precision; what is more, the redshift is not needed as an input. Nevertheless, the \( 5_{\text{max}} \) is not able to retrieve the redshift of the object when galaxies have low emission. The set of redshifts that maximizes the sum of the EWs is largely uncertain and consequently we do need the PDFs to constrain the redshift value. In order to explore the limitation of the model as a function of the redshift uncertainty and the EW of each one of the emission lines, we assemble galaxies in bins by the EW provided in the SDSS catalog and compute the ratio \( R \) between the predicted and observed EW. Each bin contains 500 galaxies in the interval \( 10^{\gamma} < \text{EW}_{\text{SDSS}} < 10^{\gamma+0.1} \) with \( \gamma \) ranging from 0.8 to 2.5 for \( \text{H}\alpha \), from 0.8 to 2.2 for \([\text{OIII}]5007\), from 0.8 to 1.8 for \( \text{H}\beta \) and from 0.8 to 1.8 for \([\text{NII}]6584\). As we observe in Fig. 9, \( \text{H}\alpha \) is clearly more affected by the \( 5_{\text{max}} \) strategy when \( \text{EW}(\text{H}\alpha) \leq 10^{-2} \, \text{Å} \). Independently of the redshift uncertainty, the \( \text{ANN}_{\alpha} \) trained with the CALMa set has more difficulties to constrain the \([\text{NII}]6584 \) line underestimating its value as the EW increases. It also presents more dispersion, which might be an indication that 500 galaxies are not enough to study this dependency properly. Nonetheless, we are able to constrain the EW of galaxies with a bias less than 10% for most of the lines even with high uncertainty in the redshift. What is more, if we compare our technique with the precision that can be obtained with traditional methods the improvement is remarkable. Let us assume, that an emission line falls within one filter and we know with high precision the redshift of the object. The EW of an emission line can be computed assuming the line is infinitely thin as:

\[
\text{EW} = \Delta'(\lambda) (Q - 1)
\]

where \( \Delta' \) is the effective width of the filter and \( Q \) is the ratio between the flux with and without emission line see (see Pascual et al. 2007, for details) or simply:

\[
Q = 10^{0.4\text{EW}_{\Lambda m}/\text{EW}_{\Lambda m}^\text{cont}}^{1/2.5}
\]

in AB magnitudes. Then, if we are able to estimate the flux of the stellar continuum in the filter tracing the emission line, obtaining the EW is straightforward. The \( S/N \) of such line depends only on \( Q \) and the precision on which the photometry can be measured \( (\Delta f(\lambda)/f(\lambda)) \) where \( f(\lambda) \) stands for the flux:

\[
\frac{\text{EW}}{\Delta \text{EW}} = \frac{Q - 1}{\ln 10 \log e[(\Delta f(\lambda)/f(\lambda))]}
\]

For \( S/N = 3 \) and an error in the photometry of 1.5%, the minimum EW measurable in a filter width of 150 Å is approximately 10 Å while our method is able to reach on average the same \( S/N \) for the same EW (upper-left panel in Fig. 9) and more importantly: with an error in the photometry of 10%. This fact illustrates once again the capability of machine learning algorithms to go beyond in precision and accuracy respect to traditional methods when large amount of data are available.

### 5. Comparison between miniJPAS and SDSS

In this section we analyze and compare the data from the SDSS survey that has also been observed with miniJPAS in the AEGIS field. Firstly, we describe the miniJPAS survey in Sect. 5.1. Then, we analyze and compare the properties of galaxies in terms of their emission lines in Sect. 5.2.

#### 5.1. miniJPAS survey

The miniJPAS survey (Bonoli et al. 2020) is the result of the J-PAS-Pathfinder observation phase carried out with the 2.55 m telescope (T250) at the Observatorio Astrofísico de Javalambre in Teruel (Spain). miniJPAS was observed with the Pathfinder camera, the first instrument installed in the T250 before the arrival of the Javalambre Panoramic Camera (JPcam. Cenarro et al. 2019; Taylor et al. 2014; Marin-Franch et al. 2015). JPAS-Pathfinder instrument is a single CCD direct imager (9.2k×9.2k, 10μm pixel) located at the center of the T250 FoV with a pixel scale of 0.23 arcsec pix\(^{-1}\), that is vignette on its periphery, providing an effective FoV of 0.27 deg\(^2\). The miniJPAS data includes four pointings of 1 deg\(^2\) in total along the Extended Groth Strip (called the AEGIS field). We use the same photometric system of J-PAS. Thus, AEGIS was observed with 56 narrow band filters covering from \( \sim 3400 \) to \( \sim 9400 \, \text{Å} \). Observations in the four broad bands (\( u_{\text{JPAS}} \), and SDSS g, r, and i) were also taken. More than 60000 objects were detected in the r band, allowing to build a complete sample of extended sources up to \( r \leq 22.7 \)
Fig. 6. BPT diagram obtained with the ANN_R and SDSS testing sample from the MPA-JHU DR8 catalog. The ANN_R is trained with the CALMa set. The color-code indicates the density of points. The solid (ka03), dashed (Ke01) and dotted lines (S07) define the regions for the four main ionization mechanism of galaxies. The percentage for each group is shown in black.

Fig. 7. BPT diagram obtained by the ANN_R trained with the CALMa set. Arrows point in the direction towards the location where galaxies should be placed according to their position in the SDSS MPA-JHU DR8 catalog. The color represents the distance for each point between the two BPT diagrams. The solid (ka03), dashed (Ke01) and dotted lines (S07) define the regions for the four main ionization mechanisms of galaxies. The percentage for each group is shown in black. The histograms on the rights represent the angular distribution of the arrows for Star forming, Seyfert and composite galaxies. The angle is defined as a clockwise rotation towards the x axis.

(AB). A detailed description of the survey is in Bonoli et al. (2020). Data is accessible and open to the community through the web page of the survey.\footnote{http://www.j-pas.org/}

5.2. miniJPAS vs SDSS

For this comparison, we select all galaxies observed with SDSS and miniJPAS with redshift below $z \leq 0.35$ and minimum average S/N of 20 in J-PAS narrow band filters. By a visual inspection we get rid of all QSOs in the sample. We end up with a total of 89 objects. Whenever photometry measurements are lacking or the S/N in a particular filter is below 2.5, we replace
Table 2. Relative difference between the EWs (in percentage) and ratios (in dex) predicted by ANN$_R$ and the values provided by the testing samples. The comparison is made between the training sample proposed in this paper and SDSS, CALIFA and MaNGA testing sample.

| Training vs Test | $H\alpha$ (%) | $H\beta$ (%) | [OIII] (%) | [NII] (%) | $[\text{OIII}]/H\alpha$ [dex] | $[\text{OIII}]/H\beta$ [dex] | O 3 N 2 [dex] |
|-----------------|--------------|--------------|------------|-----------|-----------------------------|-----------------------------|-------------|
| SDSS vs SDSS    | $-0.4 \pm 0.5$ | $-2.1 \pm 1.2$ | $1.7 \pm 1.6$ | $2.5 \pm 1.6$ | $0.014 \pm 0.092$ | $0.018 \pm 0.090$ | $0.008 \pm 0.119$ |
| SDSS vs CALIFA  | $-6.3 \pm 0.6$ | $-2.4 \pm 1.3$ | $-5.5 \pm 2.0$ | $-2.1 \pm 2.0$ | $0.018 \pm 0.121$ | $0.022 \pm 0.110$ | $0.022 \pm 0.101$ |
| SDSS vs MaNGA   | $-2.4 \pm 1.1$ | $-8.0 \pm 1.8$ | $-3.3 \pm 1.9$ | $10.0 \pm 2.1$ | $0.057 \pm 0.106$ | $0.023 \pm 0.096$ | $-0.035 \pm 0.152$ |
| CALMa vs CALIFA | $-4.2 \pm 0.5$ | $-4.8 \pm 1.2$ | $1.8 \pm 1.9$ | $-3.6 \pm 1.5$ | $0.003 \pm 0.088$ | $0.033 \pm 0.094$ | $0.034 \pm 0.113$ |
| CALMa vs MaNGA  | $-1.9 \pm 0.8$ | $-1.4 \pm 1.2$ | $0.1 \pm 1.8$ | $8.6 \pm 1.8$ | $0.045 \pm 0.085$ | $0.009 \pm 0.086$ | $-0.035 \pm 0.131$ |

The median redshift in the SDSS catalog for a total of 10000 galaxies. Points are color-coded with the spectroscopic redshift.

![Redshift retrieval](image)

**Fig. 8.** $\delta z$ obtained from the difference between the spectroscopic redshift and the median redshift in the $S_{\text{max}}$ setting in function of the sum of the EWs provided in the SDSS catalog for a total of 10000 galaxies. Points are color-coded with the spectroscopic redshift.

it by the best-fit obtained from the stellar population analysis of the galaxy as we discussed in Sec. 3.5. For this comparison we employ BaySEaGal (Amorim in prep), a Bayesian parametric approach which assumes a tau-delayed star formation model for the star formation history. Generally, galaxy properties vary within the galaxy: the gas, its temperature and its density, the distribution of interstellar dust or the stellar populations change in function of the position in the galaxy (González Delgado et al. 2015). Consequently, if the SFR of a galaxy were higher in the outer parts, the galaxy would look younger in the integrated spectrum than in the central part. Similarly, the AGN of a galaxy would not leave the same imprint in the spectrum if the integrated areas covered regions dominated by other ionization mechanisms. Therefore, ideally, one would like to analyse the same region in both surveys, which implies integrating over the same area. However, the aperture corresponding to the 3 arcsec fiber of SDSS is not sufficiently large to ensure that the Point Spread function (PSF) of J-PAS filter system is not affecting the photometry in the filters where the seeing is worse. For this reason, we make use of the MAG_PSF_COR photometry which corrects each magnitude individually by considering the light profile of the galaxy and the PSF for each filter (Molino et al. 2014, 2019). As a consequence, the integrated area varies from galaxy to galaxy, going from 2 to 7 arcsec, and should be taken into account to interpret fairly this comparison. Although the ANN$_R$ only use colors as inputs, we scale the SDSS spectrum to match the rSDSS J-PAS magnitude in each galaxy for a visual inspection.

Figure 10 shows the EWs obtained by the ANN$_R$ on J-PAS photometric data (column 1) and on the synthetic J-PAS magnitudes obtained after convolving SDSS spectra with J-PAS filters (column 2). We compare those values with the EWs derived as a result of fitting a Gaussian function to each one of the emission lines in the spectrum (x-axis). We do not include in this comparison the emission lines where EWs are below 1 Å, which indeed are compatible with zero. The number of galaxies in each row are from top to bottom 57, 37, 64, and 31. We find an excellent agreement when it comes to SDSS synthetic magnitudes, which is in line with the simulations performed with the SDSS dataset. We also find a remarkable correlation in $H\alpha$, $H\beta$ and [NII] with J-PAS magnitudes, but we obtain in most of the cases higher values with an increase in the dispersion (see median and MAD in Fig. 10). The agreement is less favourable for [OIII] line. Nevertheless, we should bear in mind the limiting number of galaxies used here in order to avoid drawing any conclusion that may not be supported from a statistical point of view. Instead, we consider more appropriated
to analyze the origin of these discrepancies by examining visually each object.

In Fig. 11 we show several galaxies analyzed in this comparison. We re-scale the SDSS spectrum to match the rSDSS J-PAS magnitude. We compare the values of the EWs measured in the SDSS spectrum (black) with the values predicted by the ANN (blue) for each one of these galaxies. On the bottom part, we show in each filter the difference between J-PAS data and SDSS synthetic photometry, which certainly can help to shed light on the origin of the discrepancies. In the first row we display three examples of emission line galaxies where the agreement in most of the EWs is remarkable. Although ANNs are often difficult to interpret, it is evident after a visual inspection that the filters capturing the fluxes of the emission lines are the most relevant in determining the values of the EWs. The excess in the flux of Hα in galaxy 2243-8838 explains the increase in its EW respect to what it is obtained from a direct measurement in the spectrum or with the synthetic fluxes by means of the ANN. In the same vein, the drop in the flux observed in the [OIII] line in galaxy 2241-12850 clarifies the differences found in the EW. Second order terms include the relation between emission lines (Balmer decrement or recombination lines) and the colors of galaxies. Definitely, the excess in the flux of Hβ in galaxy 2243-9127 does not only increase the value of such line but also contributes to enlarge the EW of Hα. In the second row of Fig. 11 we show Early-type galaxies (ETGs) where the differences between J-PAS data and SDSS synthetic fluxes are negligible. The ANN estimates very low probability for these galaxies to have any emission line with a EW greater than 3 Å, which is in agreement with the measurements performed in SDSS spectra. As we discussed in Sect. 4.4 the ANN tends to overestimate the EWs in the regime of low emission and consequently a zero level bias appears in these galaxies. Nonetheless, for many of these lines the values are compatible with the uncertainty and never overcome the 3 Å limit. Finally, in the third row of Fig. 11 we focus our attention on galaxies where the fluxes seen by J-PAS and SDSS present evident differences in the blue part of the spectrum. The integrated areas in J-PAS are probably capturing regions with more populations of young stars in 2243-9209 and 2406-4867 galaxies. Such population rises the number of ionising photons and it is responsible of the increase in the EWs of emission lines that we observe. The opposite effect occurs in galaxy 2406-5886, the galaxy looks redder with J-PAS data and the flux in Hα is less intense. Therefore, the predictions of the ANN in the EWs are below the values measured in the SDSS spectrum.

To sum up, despite of the fact that this comparison suffer from several difficulties and it would need many more galaxies to be statistically robust, results are coherent with the simulations presented in Sect. 4 and lay the foundations to better understand and interpret the whole sample of galaxies observed in the AEGIS field that we will analyze in a future work.

6. Summary and conclusions

We have developed a new method based on ANNs to measure and detect emission lines in J-PAS up to $z = 0.35$. We can classify galaxies according to the EWs of the emission lines even with high uncertainty in the redshift. This will allow us to better study the density function of emitting-line galaxies in J-PAS. With the synthetic photometry of CALIFA, MaNGA or SDSS spectra, the ANNs can be trained to estimate the EWs of Hα, Hβ, [NII]λ6584 and [OIII]λ5007 lines. We present two training samples to undertake this task. Firstly, we trained the ANN with only synthetic J-spectra from MaNGA and CALIFA surveys and we used SDSS to evaluate the performance of the model. The lack of enough number of AGN-like synthetic J-spectra produces a saturation of $[NII]/Hα$ and $[OIII]/Hβ$ ratios at high values, which compromises the ability of the model to deal with galaxies where the main ionization mechanism is not dominated by star formation processes. Nevertheless, we are able to constrain those ratios within 0.101 and 0.091 dex. Furthermore, we are able to reach 0.091 and 0.087 dex respectively.
Fig. 11. Examples of J-PAS galaxies in the AEGIS field with SDSS spectrum. The SDSS spectrum is re-scaled to match the rSDSS J-PAS magnitude. Diamonds correspond to the filters not used by the ANN. Blue and black numbers show, respectively, the predictions made by the ANN and J-PAS data were estimated to reach a precision of 0.16 dex in [NII]/Hα (Benitez et al. 2014). Secondly, we trained the ANN with SDSS galaxies and we revealed the importance of testing the model with data coming from different surveys. Otherwise, the performance of the model can be overestimated.

While the SDSS training set scores very high with SDSS testing sample, the performance worsens when we compare it with MaNGA or CALIFA test sample.

Finally, we estimate the EWs of a set of galaxies observed in different surveys in equal terms, we reach an overall

if one considers only star-forming galaxies. This is a significant improvement in the precision previously expected. Methods based on the similarities between synthetic J-spectra from SDSS or other surveys with accessible information to emission line and J-PAS data were estimated to reach a precision of 0.16 dex in [NII]/Hα (Benitez et al. 2014). Secondly, we trained the ANN with SDSS galaxies and we revealed the importance of testing the model with data coming from different surveys. Otherwise, the performance of the model can be overestimated.

While the SDSS training set scores very high with SDSS testing sample, the performance worsens when we compare it with MaNGA or CALIFA test sample.

Finally, we estimate the EWs of a set of galaxies observed both in SDSS and miniJ-PAS. We compare the performance of ANN in the synthetic SDSS fluxes with the performance in the fluxes measured by J-PAS. Despite the difficulty of comparing data from different surveys in equal terms, we reach an overall
agreement. We argue that the origin of the discrepancies might be attributed to differences between the integration areas in miniJPAS and SDSS and/or photometry artefacts that appear as a result of the PSF. Many more data would be needed to be conclusive. In this work our model is limited to redshift below $z = 0.35$ in order to ensure $H\alpha$ line is measurable with the J-PAS filter system. However, J-PAS will be able to detect galaxies up to $z \sim 1$. Other emission lines such as the $[OII] \lambda \lambda 3726,3729$ doublet are visible in the optical range up to redshift $z < 1.6$ and have been used as tracer of star formation in many works (Kewley et al. 2004; Sobral et al. 2012). An ultimate version of the model should take into account those facts and build a more sophisticated and complete training sample to be able to overcome the limitations and inaccuracies mentioned so as to fully exploit the potentiality of J-PAS. Our main conclusions are summarized below:

- The ANN$_C$ can classify galaxies according to the EWs of the emission lines beyond the contrast that one can directly measure with sufficient significance in J-PAS (~ 10 Å) and also in the case of high uncertainty in the redshift.
- The ANN$_R$ trained with the CALMa set can estimate the EWs of $H\alpha$, $H\beta$, $[NII]$, $[OIII]$, $[OII]$, $[OIII]_\lambda 5007$ in SDSS galaxies with a relative standard deviation of 8.7%, 14.3%, 15.9%, and 16.4% respectively, $H\alpha$, $H\beta$, $[NII]$, $[OIII]$, and $[OIII]_\lambda 5007$ lines presents a relative bias of 0.17%, 5.4%, 4.8%, and −6.4% respectively.
- The $[NII]/H\alpha$ is constrained within 0.10 dex and a bias of −0.028 dex and the $[OIII]/H\beta$ ratio with no bias and a dispersion of 0.091 dex in SDSS galaxies. The $O\ 3N\ 2$ is recovered within 0.112 dex and a bias of 0.038 dex.
- We found an overall correlation between miniJPAS and SDSS galaxies in the EW of $H\alpha$, $H\beta$, and $[NII]$. The correlation in the EW of $[OIII]_\lambda 5007$ is less strong. More data will be needed to unveil the origin of such discrepancy. Certainly, the problems associated to the integrated areas are playing an important role.

Acknowledgements. G.M.S., R.G.D, R.G.B., E.P., J.R.M, and J.M.V. acknowledge support from the State Agency for Research of the Spanish MCIU through the “Center of Excellence Severo Ochoa” award to the Instituto de Astrofísica de Andalucía (SEV-2017-0709) and the projects AYA2016-77846-P and PID2019-109067-GB100. L. A. D. G. acknowledges support from the Ministry of Science and Technology of Taiwan (grant MOST 106-2628-M-001-003-MY3) and from Academia Sinica (grant AS-IA-107-M01). P.O. Baqui and V. Marra acknowledge support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. R.A.D. acknowledges support from the CNPq through BP grant 308105/2018-4, and FINEP grants Ref. 1217/13 - 01.13.0179/00 and Ref 0859/10 - 01.10.0663/00 and also FAPERJ PRONEX grant E-26/110.566/2010 for hardware funding support for the J-PAS project through the National Observatory of Brazil and Centro Brasileiro de Pesquisas Físicas. LSJ acknowledges support from Brazilian agencies FAPESP (2019/10923-5) and CNPq (304819/2019-94), AC acknowledges financial support from research projects AYA2016-79724-C4-4-P, PID2019-107408GB-C44 from the Spanish Ministerio de Ciencia e Innovación. The authors acknowledge the following people for providing valuable comments and suggestions on the first draft of this paper: Stravos Akras, Joel Bregman, Salvador Duarte Puertas, Jorge Iglesias, Yolanda Jimenez Teja, Jose Miguel Rodriguez Espinosa, David Sobral and, Adi Zitrin. This research made use of Python (http://www.python.org), Numpy (Van Der Walt et al. 2011), Matplotlib (Hunter 2007), a suite of open-source Python modules that provides a framework for creating scientific plots and, Astropy, the community-developed core python package (Astropy Collaboration et al. 2013; Price-Whelan et al. 2018). Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org. This study makes use of the results based on the Calar Alto Legacy Integral Field Area (CALIFA) survey (http://califa.caha.es/). This project made use of the MaNGA-Pipe3D dataproducts. We thank the IA-UNAM MaNGA team for creating this catalogue, and the
Appendix A: SDSS training set

In this section we show how the SDSS training set scores in the SDSS testing sample. This represents the ideal situation where the testing set is included within the parameter space of the training set. In other words, the testing sample is a subset of the training set and consequently the only uncertainties found in the targets variables (EWs) area associated to the capability of the ANN$_R$ to predict in J-PAS data. As we discussed in the main body of this paper, here lies the reason why the ANN$_R$ must be tested with data with different observational setup and calibrations.

In Fig. A.1 we plot the EWs predicted by the ANN$_R$ versus the EWs provided by the SDSS testing sample from the MPA-JHU DR8 catalog. This plot follows the same scheme of Fig. 4. As happened with the CALMa training set, we constrain the EW of $H_{\alpha}$ followed by [OIII]$_{\lambda5007}$ and [NII]$_{\lambda6584}$. However, the [NII]$_{\lambda6584}$ line is recovered with no bias and it does not saturate at high values.

In Fig. A.2 we show the comparison between the logarithmic ratios of [NII]/$H_{\alpha}$, [OIII]/$H_{\beta}$ and O3N2 in a similar way as we did in Fig. 5. The [NII]/$H_{\alpha}$ ratio is predicted within 0.092 dex and a bias of 0.014 dex and the [OIII]/$H_{\beta}$ ratio within 0.090 dex and a bias of 0.018 dex. As a result, the O3N2 is recovered within 0.119 dex and no bias.

Finally, we show in Fig. A.3 a comparison of the BPT diagram recovered by the ANN$_R$ (left plot) and the one obtained from the SDSS testing sample (right plot) following once again the same scheme of Fig. 6. The similarity between those diagrams is remarkable. We are not only able to recover properly the SF-wing but also the AGN branch, obtaining similar percentages of galaxies in all the regions.

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from tensorflow.org

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33, [1307.6212]

Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, PASP, 93, 5.

Bamford, S. P., Rojas, A. L., Nichol, R. C., et al. 2008, Monthly Notices of the Royal Astronomical Society, 391, 607, [http://oup.prod.sis.lan/mnras/article-pdf/391/2/5007/5707752/mnras0391-0607.pdf].

Baqui, P. O., Marra, V., Casarini, L., et al. 2020, arXiv e-prints, arXiv:2007.06222.

Benitez, N., Dukpe, R., Moles, M., et al. 2014, arXiv e-prints, arXiv:1403.5237,

Bonneau, V., Aghanim, N., Salomé, P., et al. 2019, A&A, 622, A137, [1901.01932]

Bonoli, S., Marin-Franch, A., Varela, J., et al. 2020, arXiv e-prints, arXiv:2007.01910, [2007.01910].

Brinchmann, J., Charlton, S., White, S. D. M., et al. 2004, MNRAS, 351, 1151, [astro-ph/0310660].

Bundy, K. 2015, in IAU Symposium, Vol. 311, Galaxy Masses as Constraints of Formation Models, ed. M. Cappellari & S. Courteau, 100–103

Cañada-Torrecilla, C., Gil de Paz, A., Castillo-Morales, A., et al. 2015, A&A, 576, A135, [1409.8302].

Gomes, J. M., Papaderos, P., Vílchez, J. M., et al. 2016, A&A, 586, A22, [1510.01900].

González Delgado, R. M., García-Benito, R., Pérez, E., et al. 2015, A&A, 581, A103, [1506.04157].

Hunter, J. D. 2007, Computing in Science & Engineering, 9, 90.

Kauffmann, G., Heckman, T. M., Tremonti, C., et al. 2003a, MNRAS, 346, 1055, [astro-ph/0303239].

Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003b, MNRAS, 341, 33, [astro-ph/0204055].

Kelz, A., Verheijen, M. A. W., Roth, M. M., et al. 2006, PASP, 118, [astro-ph/0512557].

Kocevski, D. L., Dopta, M. A., Sutherland, R. S., Heisler, C. A., & Trevena, J. 2001, ApJ, 556, 121, [astro-ph/0106324].

Kewley, L. J., Geller, M. J., & Jansen, R. A. 2004, AJ, 127, 2002, [astro-ph/0401172].

Kewley, L. J., Nicholls, D. C., & Sutherland, R. S. 2019, A&A, 57, 511, [1910.09730].

Lacerda, E. A. D., Cid Fernandes, R., Couto, G. S., et al. 2018, MNRAS, 474, 3727, [1711.07844].

Madau, P. & Dickinson, M. 2014, A&A, 52, 415, [1403.0007].

Masuñón, R. & Mannucci, F. 2019, A&A Rev., 27, 3, [1811.09642].

Marin-Franch, A., Taylor, K., Cenarro, J., Cristóbal-Hornillos, D., & Moles, M. 2015, in IAU General Assembly, Vol. 29, 2257381

Molino, A., Benítez, N., Moles, M., et al. 2014, MNRAS, 441, 2891, [1306.4968].

Moreno, A. Costa-Duarte, M. V., Mendes de Oliveira, C., et al. 2019, A&A, 622, A178, [1804.03640].

Nair, V. & Hinton, G. E. 2010, in Proceedings of the 27th International Conference on International Conference on Machine Learning. ICML’10 (Madison, WI, USA: Omnipress), 807–814

Pascale, S., Gallego, J., & Zamorano, J. 2007, PASP, 119, 30, [astro-ph/0611121].

Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, A&A, 621, A26, [1806.06607].

Pettini, M. & Pagel, B. E. J. 2004, MNRAS, 348, L59, [astro-ph/0401129].

Price-Whelan, A. M., Sipőcz, B. M., Günther, H. M., et al. 2018, AJ, 156, 123.

Roth, M. M., Kelz, A., Fechner, T., et al. 2005, PASP, 117, 620, [astro-ph/0502581].

Sánchez, S. F., Kennicutt, R. C., Gil de Paz, A., et al. 2012, A&A, 538, A8, [1111.0962].

Sánchez, S. F., Pérez, E., Sánchez-Blázquez, P., et al. 2016a, Rev. Mexicana Astron. Astrofis., 52, 171, [1602.01830].

Sánchez, S. F., Pérez, E., Sánchez-Blázquez, P., et al. 2016b, Rev. Mexicana Astron. Astrofis., 52, 21, [1509.08352].

Schawinski, K., Thomas, D., Sarzi, M., et al. 2007, MNRAS, 382, 1415, [0709.3015].

Sobral, D., Bendo, G. J., & Muzzin, A. 2012, MNRAS, 420, 1926, [1109.1830].

Taylor, K., Marin-Franch, A., Laporte, J. R., et al. 2014, Journal of Astronomical Instrumentation, 3, 1350010, [1301.4715].

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science & Engineering, 13, 22.

Villada-Rojo, G., Vironen, K., López-Sanjuan, C., et al. 2015, A&A, 580, A47, [1505.07115].

Whitton, D. D., Placco, V. M., Beers, T. C., et al. 2019, A&A, 622, A182, [1811.02279].

York, D. G., Adelman, J., Anderson, J. E., et al. 2000, AJ, 120, 1579, [astro-ph/0006396].
Fig. A.1. EWs of $H\alpha$, $H\beta$, [NII]λ6584 and [OIII]λ5007 predicted by the ANN compared to SDSS testing sample. The ANN is trained with the SDSS training set. The color-code represents the probability density function defined by a Gaussian kernel (right panel) and the redshift of the objects (left panel). The histograms are normalized to one and show the relative difference between both values. Black and blue numbers are the median and the median absolute deviation of the difference. Black and grey dashed lines on the left are lines with slope one and the best linear fit respectively. We perform a sigma clipping fit with $\sigma=3$ to exclude outliers. Red dashed line represents the median.
Fig. A.2. Comparison between \([\text{NII}]/H\alpha\), \([\text{OIII}]/H\beta\) and \(\text{O3N2}\) ratios estimated by the ANN\(_R\) and SDSS testing sample. The ANN\(_R\) is trained with the SDSS training set. Same scheme of Fig. A.1.
Fig. A.3. BPT diagram obtained with the ANN_{P} and SDSS MPA-JHU DR8 catalog where the color-code indicates the density of points. The ANN_{P} is trained with the SDSS training set. The solid (ka03), dashed (Ke01) and dotted lines (S07) define the regions for the four main ionization mechanism of galaxies. The percentage for each group is shown in black.