Mon. Not. R. Astron. Soc. **000**, 1–?? (2015)   Printed 23 March 2016   (MN LaTeX style file v2.2)

The Machine: An Artificial Neural Network to Understand the Complexities of Emission Line Profiles

E.J. Hampton,[1]⋆ A. Medling[1], B. Groves[1], R. Davies[1], M. Dopita[1], I-T. Ho[11], M. Kaasinen[1], L. Kewley[1], S. Leslie[1], R. Sharp[1], S.M. Sweet[1], A.D. Thomas[1], SAMI team people, S7 team people

[1] *Research School of Astronomy & Astrophysics, Australian National University, Canberra, ACT 2611, Australia*
[11] *Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, HI 96822, USA*

Accepted . Received ; in original form

ABSTRACT

Integral Field Spectroscopy (IFS) surveys are changing how we study galaxies and creating more data than we have had before. The large number of resulting spectra makes emission line fitting with visual inspection an unfeasible option. Here we present `The Machine`, an artificial neural network (ANN) that determines the number of Gaussian components needed to describe the complex emission line velocity structures observed in galaxies. To demonstrate `The Machine's` capabilities we have applied it to two distinct surveys using two different IFS instruments; the S7 survey using the Wide Field Spectrograph and the SAMI galaxy survey. We demonstrate that using an ANN is comparable with astronomers in determining the best number of Gaussian components to describe the physical processes in galaxies. The advantage of our ANN, `The Machine`, is that is capable of processing the spectra of thousands of galaxies in minutes compared to the years it would take individual astronomers to complete the same task by visual inspection.

Key words: Emission line:galaxies – neural network – multicomponent fitting – optical:galaxies

1 INTRODUCTION

Integral Field Spectroscopy (IFS) is changing our approach to studying galaxy evolution. Surveys such as CALIFA (Calar Alto Legacy Integral Field Area, Sánchez et al. 2012), SAMI (Sydney-AAO Multi-object Integral Field, Croom et al. 2012), MaNGA (Mapping Nearby Galaxies at Apache Point Observatory, Bundy et al. 2015), and S7 (Siding Spring Southern Seyfert Spectroscopic Snapshot Survey, Dopita et al. 2014) are building databases of spatially resolved spectra of hundreds to thousands of galaxies in order to explore galaxy evolution as a function of morphological and spectroscopic classification, and environment. Integral Field Spectroscopy provides a powerful probe into the spatial variation of physical processes across galaxies. For example, single-fibre redshift surveys such as SDSS (Sloan Digital Sky Survey) and GAMA (Galaxy And Mass Assembly, Driver et al. 2009) observe only a single spectrum for each galaxy, typically from its core. As such, one can often misidentify the global properties of a galaxy, (e.g. Richards et al. 2014; Ho et al. 2014; Hampton et al. in prep).

Figure 1 highlights the situation of changing physical processes across a galaxy showing the variation in emission line profiles between the core (orange) and away from the core (magenta) of the S7 galaxy NGC 5728. The orange lines show an obvious double peaked emission line profile attributed to gas moving in two directions, e.g. an outflow of gas in the disk of the galaxy. IFS provides a powerful probe to understand the broader nature of galaxies.

Access to the wealth of information from an IFS survey comes at a price: data volume. Not only are advances in IFS technology pushing the previous sample size boundaries but each galaxy observation now contains as many individual spectra as an entire early redshift survey. Datacubes of multiple gigabytes, with thousands of spaxels (spatial pixels) for each galaxy, are not uncommon.

Data reduction pipelines (e.g. Husemann et al. 2013; Sharp et al. 2015; Allen et al. 2015a) are efficient ways to convert raw data into a final spectral datacube for analysis, but interpreting these spectra remains a significant challenge. The data volume is too great to allow visual inspection

⋆ E-mail:elise.hampton@anu.edu.au

of all spectra individually. Some form of automated analysis is required to extract the important information from the spectra and to target galaxies for further investigation.

Automated continuum and absorption line fitting is routinely used to understand the stellar populations within galaxies, subsequent emission line fitting provides insight into active star formation, AGN (active galactic nuclei) activity and shock properties of galaxies. This type of preanalysis is time consuming for IFU surveys and fitting each emission line by hand is no longer a feasible option. We now understand that there can be multiple processes behind a single emission line, creating further steps to our preanalysis. Automated emission line fitting, including multicomponent fitting for situations with multiple physical processes contributing to emission lines, are currently in use (e.g. `LZIFU`; Ho et al. in prep). However there is still human input required to decide if more than one component is necessary to describe each emission line. This paper describes our automated machine learning algorithm to remove this time consuming human input and streamline multicomponent emission line fitting for large surveys.

2 SPECTRAL PROPERTIES FROM DATA CUBES

Analysis of a galaxy data cube requires the measurement of key physical properties extracted from individual spectra of each spaxel (spatial pixel). The spectrum at each spaxel typically contains an emission-line spectrum, arising in shock-heated or photoionised gas, superimposed upon continuum light, either from the underlying stellar populations or an active galactic nucleus. Accurate modelling and subtraction of the underlying continuum is critical in correcting for stellar absorption which would otherwise lead to the incorrect measurement of coincident emission-lines, e.g. predominantly those of the Hydrogen Balmer series, Hα and Hβ, in which stellar absorption can have a high equivalent width.

Each line of sight into the galaxy can encompass gas at different velocities and with different excitation mechanisms. We fit multiple Gaussians to the resulting composite emission-lines in order to explain the underlying physical processes occurring within a galaxy. However, we cannot know the number of physical components within a resolution element.

To fit the spectral data cubes we use the automated fitting package LZIFU (Ho et al. in prep). This program, written in the IDL programming language, fits multiple Gaussian components to each emission-line complex in a spectrum after correcting for the underlying stellar absorption component. The emission lines are fit simultaneously, and thus each component has a single velocity and velocity dispersion. However, the relative fluxes of the emission lines are left free, and the line ratios for each component can vary (for a full description see Ho et al. in prep).

S7 (The Siding Spring Southern Seyfert Spectroscopic Snapshot Survey; Dopita et al. 2014) and SAMI (the SAMI galaxy survey; Croom et al. 2012) are the first two galaxy surveys we have used in our study to determine if machine learning can help with the time constraints of getting data out for a large survey. For each galaxy observed we have ~1000 spaxels with each spaxel associated with a high spectral resolution spectrum. Each spectrum is fit in turn with 1, 2, and 3 Gaussian components for the strongest emission lines. The significant challenge is in identifying which set of Gaussian components best describes the data at each spaxel. Visual inspection to make the identifications, which is the common approach, for a single galaxy can take up to 1 hour. For small surveys this is feasible, but with surveys the size of SAMI (~3000 galaxies) this means ~ 125 days continuous work, or multiple years for a single astronomer.

F-tests are used to automatically identify the best fit number of Gaussians to spectral emission lines, by determining if increasing the number of gaussians increases the significance of the fit. However, the f-test is based on the χ^2 value of the fit which, when using multiple Gaussians, may fall into a local minimum rather than finding the global minimum as is done when using Bayesian statistics. We have also looked at the precision of using f-tests, as is done in McElroy et al. (2015), in comparison to astronomers and our machine learning algorithm, see section 6.1. Analysis has been done using Bayesian statistics on some SAMI galaxies. The results were good in determining the number of components however it took over one minute for an individual spaxel. This would mean the SAMI survey would take over 2,000 days to complete, much longer than using visual inspection.

The Machine Learning algorithm we have chosen to implement is an artificial neural network (ANN) designed to learn and make classification decisions across an entire survey, fast and reliably. By using an ANN trained by astronomers we have a system that is not only self-consistent but is able to reliably identify differences between the number of fit components and the best solution for an individual spectrum in a timely fashion.

3 A SUPERVISED ARTIFICIAL NEURAL NETWORK

For large surveys we require a reliable, self-consistent, reproducible and quick automated process of determining the best number of Gaussian components needed for each spaxel of a galaxy. We have made comparisons between individual spectroscopists and found a range of decisions for the numbers of components required for each spaxel for the same sets of galaxies. Of more concern is that repeat classification of the same galaxy by the same individual can also result in a range of choices.

The use of machine learning in Astronomy is not a new idea. Contemporary examples include the prediction of solar flares (e.g. Bobra & Couvidat 2015), understanding Gamma-ray emission from AGNs (e.g. Doert & Errando 2014; Hassan et al. 2013), and the classification of galaxy types (e.g. Kuminski et al. 2014). Machine learning covers a wide range of distinct classes of artificial intelligence (AI) such as Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), and Random Forest algorithms, that learn without being explicitly programmed. Each have their benefits and weaknesses but all are based on the same underlying principle, they learn from a training set and create models to be used to predict outcomes. We have chosen to use an Artificial Neural Network (ANN) to build a classification model for multi-component emission line fitting.

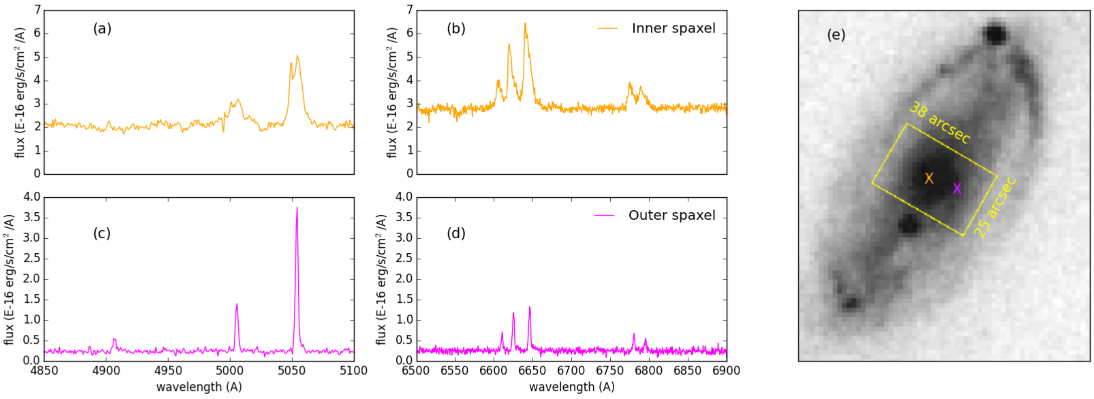

Figure 1. NGC5728. (a) Blue and (b) red spectra from a core spaxel. Note the double peak in the emission lines. (c) Blue and (d) red spectra from a spaxel away from the core of the galaxy. Note the difference in the shapes of the emission lines in comparison to the top plots. (e): DSS image of NGC5728 with WiFeS FoV overlaid in the yellow box. Magenta 'X' indicates the outer spaxel, orange 'X' indicating core spaxel.

An Artificial Neural Network (ANN) is a computer system comprised of nodes, or units, which perform calculations with a pre-determined equation. These nodes sit in layers which have different jobs depending on where they sit in the ANN design. Figure 2 presents a simplified example of an ANN to classify something as 'a cat' or 'not a cat' based on observable properties such as the 'number of legs' and 'size'. Each node in the ANN is represented by a circle. We are looking at a supervised ANN that is trained by using labelled examples, i.e. we give it answers or labels for each example to compare itself to.

An ANN has three types of layers; an input layer, hidden layers, and an output layer. Each node in the input layer, represented as x_j in the following equation, is a parameter value making up a feature vector which, in this case, describes the number of legs and size of the things we want to classify as 'a cat' or 'not a cat'.

From the input layer these two parameters are sent into the next layer, the first hidden layer of our ANN. The parameter values are put into a sigmoid function with different weights on each parameter. For each node in the first hidden layer the node performs the calculation described in equation 1 where θ_{ij} are the weights for the node i on the input parameters x_j between the input and first hidden layer.

$$a_i = \frac{1}{1 + exp(\sum_{j=1}^{2} \theta_{ij}^{1} x_j)} \tag{1}$$

Each node in the first layer uses the same parameters x_1 and x_2 but different weights θ_{ij}^1 corresponding to the specific node.

Once the values of the sigmoid functions are calculated for each node in the first layer they are passed onto the next layer, in this case the second hidden layer. The process is repeated but using the values calculated from the previous layer and again different weights corresponding to the different nodes in the second hidden layer. Equation 2 shows the functional form of the equation calculated in the nodes of the second hidden layer.

$$b_i = \frac{1}{1 + exp(\sum_{j=1}^{3} \theta_{ij}^{2} a_j)} \tag{2}$$

At this point we have reached the end of the hidden layers.

The output layer is the layer that determines the classification of our input parameters as a cat or not a cat. The values b_i from the second hidden layer are sent into the output layer where one last set of sigmoid functions are calculated with again different weights. Equation 3 shows the equation calculated by the nodes in the output layer.

$$c_i = \frac{1}{1 + exp(\sum_{j=1}^{3} \theta_{ij}^{3} b_j)} \tag{3}$$

The classification decided on by the ANN is determined by which output node has the higher value.

During the training phase of the ANN a cost is also calculated. The cost function, equation 4, describes how close the classification from the ANN was to the labels we gave it. The cost is summed over all output nodes and all training examples. The second term in the cost function sums the weights from each layer with a regularisation parameter, λ, which helps stop any particular weights from taking over the function. λ is also know as a tuneable parameter. By changing the value of lambda and comparing the results of the cost function during training we can determine the best value, between 0.01 to 10, to minimise the cost function. m is the number of training examples.

$$J = \frac{1}{m}\sum_{i}(-y_i log(c_i) - (1-y_i) log(1-c_i)) + \frac{\lambda}{2m}(\sum_{ijk}\theta_{ij}^{k}) \tag{4}$$

The cost function is then minimised in the training phase by iterating over the entire training set using an octave[1] script `fmincg`[2] The minimisation uses the cost func-

[1] `https://www.gnu.org/software/octave/`

[2] Originally written by Carl Edward Rasmussen and added to by the Stanford Machine Learning online course. `fmincg` is based on Polack-Ribiere minimisation.

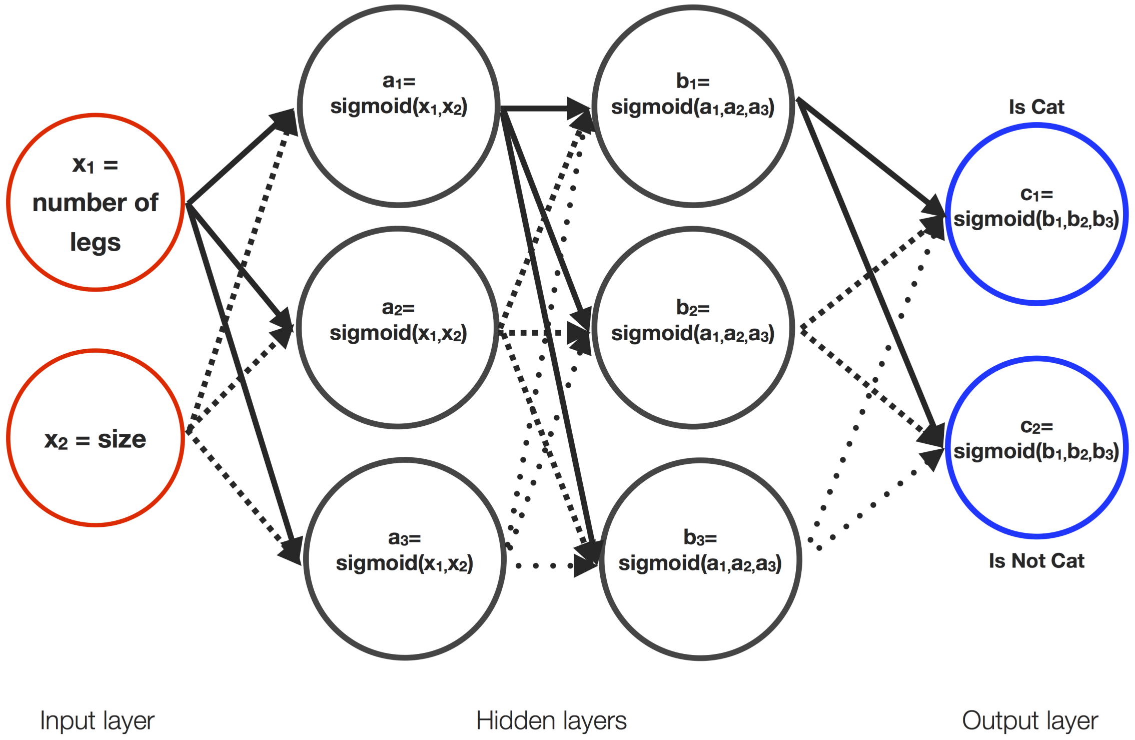
x_1 = number of legs
x_2 = size
a_1= sigmoid(x_1,x_2)
a_2= sigmoid(x_1,x_2)
a_3= sigmoid(x_1,x_2)
b_1= sigmoid(a_1,a_2,a_3)
b_2= sigmoid(a_1,a_2,a_3)
b_3= sigmoid(a_1,a_2,a_3)
Is Cat
c_1= sigmoid(b_1,b_2,b_3)
c_2= sigmoid(b_1,b_2,b_3)
Is Not Cat
Input layer
Hidden layers
Output layer

Figure 2. A simple Artificial Neural Network design. The circles indicate the nodes and are grouped into the 3 layer types: Input, hidden and output. This ANN is used to decide if something with the input parameters of size and number of legs is ‘a cat‘ or ‘not a cat‘. At each layer, calculations are done to the values of the previous layer. The final classification is then decided based on which node in the output layer has the largest value.

tion, equation 4, to alter the weights at each node to return a classification closer to the labels for each training example in the next iteration. Each successive iteration adjusts the weights again to create a decision matrix capable of matching the classifications of the labelled training examples.

In the case of our example, in figure 2, we chose two parameters but these are not enough to adequately classify something as a cat or not. For example, a small dog would be classified as a cat using this ANN. For this reason we have to give an ANN enough information to adequately describe the different classifications we would like it to make.

Our ANN, which we have called `The Machine`[3], has two hidden layers with 15 nodes in each layer. The input layer has 86 input parameters making up the feature vector for each example and the output layer has 3 nodes corresponding to the best number of components; 1, 2, or 3 components.

[3] The name of our ANN has been based on the Artificial Intelligence built by a Mr Finch in the TV series ’Person of Interest’. The outer program that controls the input and output of `The Machine` is called `Finch`.

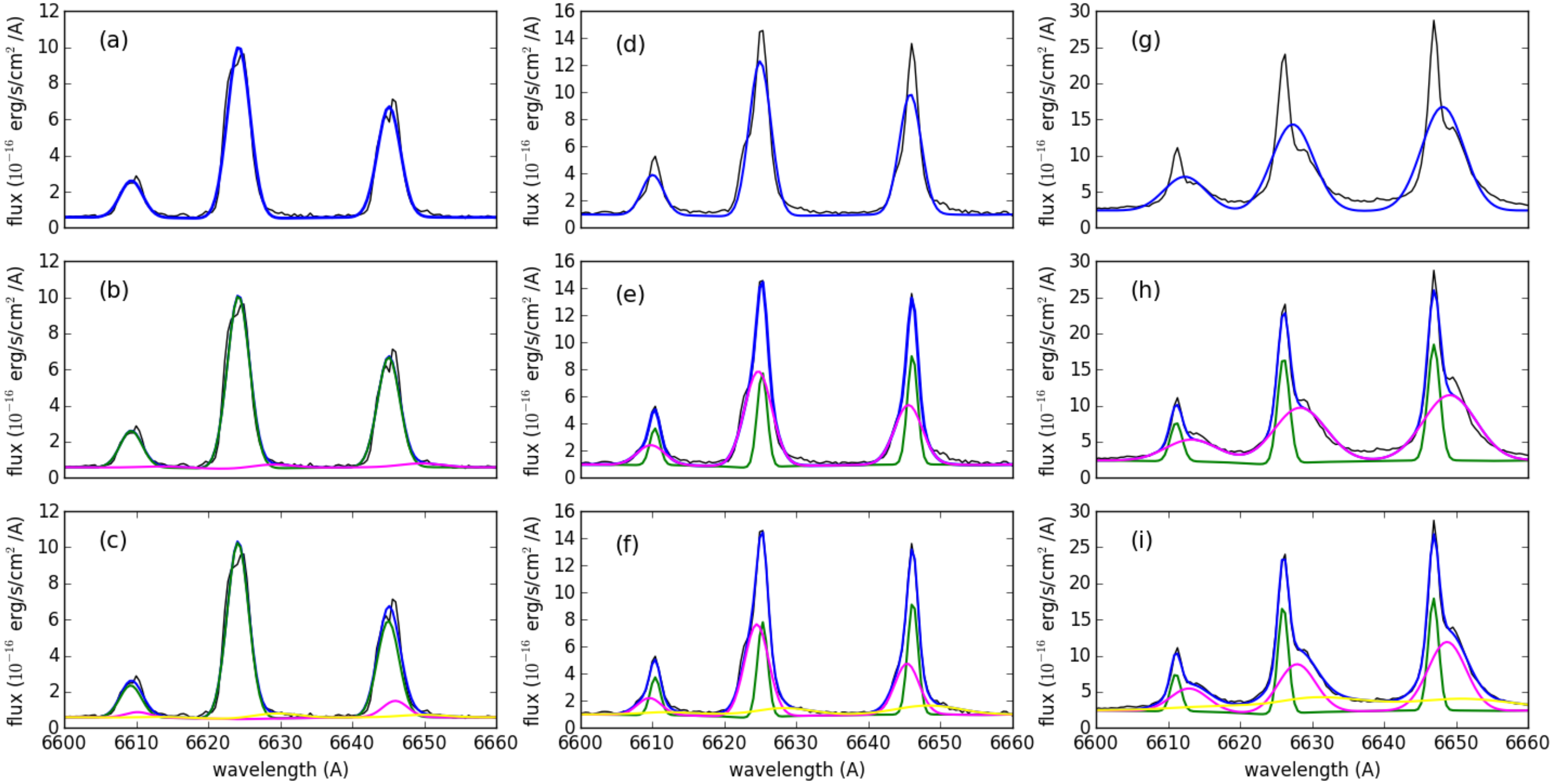

Figure 3. Each column shows an individual spectrum from a different position within the galaxy NGC5728 highlighting the LZIFU fits to [NII]$\lambda\lambda$6548, 6584 and Hα emission lines where we have high S/N and high spectral resolution spectra from S7, while each row shows the LZIFU 1, 2, and 3 component fits to these spectra, respectively. Black shows the data, blue the total line fit, green the 1st component fit, magenta the 2nd component, and yellow the 3rd component.

4 IDENTIFYING THE CORRECT MODEL

The goal of our research is to create a reliable, fast, self-consistent and easy to use method of determining the most likely number of components needed to describe an emission line. We have done this by designing an ANN to select the best number of Gaussians components for the spectra in each spaxel. As with all machine learning algorithms, we need to train the algorithm before setting it loose on survey data. The training involves giving `The Machine` labelled examples of what it might receive. To do this we use astronomers to make the decisions (section 5), labelling each spaxel with the number of components that most likely describe the emission. Then `The Machine` is trained with these examples and tested to confirm that we have set the tuneable parameters, e.g. λ, the number of iterations, and the number of training examples, are set correctly.

Our approach to training and using our ANN is as follows:

(i) Create a training set of examples of each type of emission line fit we expect to see in a survey, with labels.

(ii) Create feature vectors, or labelled examples, a numerical set of parameters associated with each example

(iii) Use half of these examples to train the ANN to build a model, using the labels to correct the weights. This will be the training set.

(iv) Use a quarter of the labelled examples as a subset to optimise the tuneable parameters of `The Machine`, not used in training. This will be the cross-validation set.

(v) Compare results of the ANN in training to the remaining quarter of labelled examples, also unseen in training. This will be the test set.

(vi) Compare the results of the ANN to each individual trainer.

The test set allows us to calculate an accuracy of `The Machine` and understand how many more examples we may need in training, while the cross-validation set allows us to tune the regularisation parameter, λ, to best suit the problem. The testing and cross-validation sets also allow us to understand how many nodes each layer should have to optimise the algorithm and how many layers the ANN needs to give the outcomes that match our human trainers closest.

Each example has 86 parameters that describe the emission line fits. The parameters were determined to be the numbers related to a fit of multiple gaussians that we believe to be important in determining if a particular fit is better than another. These cover signal-to-noise ratio of the strongest emission lines, relative contribution of each component to the total flux of an emission line, and velocity dispersion. We have used the emission line values for lines that are strong and/or commonly fit for emission line studies. All 86 input parameters making up the input feature vector into `The Machine` are listed in table 1.

5 TEST SAMPLES

`The Machine`, our ANN, is a supervised learning algorithm. The supervision comes from training with labelled examples, i.e. we give `The Machine` the answers in order to compare its classifications to. `The Machine` then uses this information to correct itself. The following subsections explain how we have used two test cases, the S7 and SAMI galaxy surveys, to test, train and run `The Machine` in order to quickly and reliably classify the number of components needed for each spaxel of a galaxy.

During the testing of the ANN we discovered that each survey has to have its own training set, due to the differences in each survey, e.g. signal-to-noise of the most common emission lines, spectral resolution, and the overall galaxies targeted by each survey (S7 has targeted galaxies with very strong Active Galactic Nuclei-like emission lines). When running the ANN on the SAMI survey after being trained with S7 the results showed no correlation with our SAMI trainers.

5.1 Siding Springs Southern Seyfert Spectroscopic Snapshot Survey : S7

The Siding Springs Southern Seyfert Spectroscopic Snapshot Survey (S7, Dopita et al. 2015) is a survey of ∼130 Seyfert galaxies, observed with the Wide Field Spectrograph (WiFeS, Dopita et al. 2010) instrument on the ANU 2.3m telescope at Siding Spring Observatory. These galaxies are at redshifts less than 0.05 and thus use most of the field of view of the WiFeS detector. S7 is intended to explore the narrow and broad line regions in Seyfert galaxies and hence has a large number of galaxies with underlying broad emission lines. For full details about S7 we refer the reader to Dopita et al. (2015). Our S7 training set is eight galaxies from the initial data release. These galaxies were chosen to cover the full range of activity within the sample; Seyfert 1's, Seyfert 2's, LINERs and star-forming galaxies.

Manual classification entails an astronomer looking at every observed spectrum along with the LZIFU fits with 1, 2, and 3 Gaussian components and the fit residuals. Concentrating on the strong emission lines, the astronomer decides the minimum number of components needed to reproduce the spectrum within the noise. As this was done for every spaxel, this resulted in a 2D component mask of the galaxy with values of 1, 2, or 3.

We found that the astronomer did not agree for $\sim 25\%$ of cases. To counteract this we trained and tested the ANN using a clean sample of ~ 2500 spaxels for which all three astronomers agreed on the number of components.

5.2 Sydney-AAO Multi-object Integral field: SAMI

The SAMI Galaxy Survey (Croom et al. 2012) is a survey of ~ 3400 nearby ($z < 0.05$) galaxies all observed with the SAMI instrument on the 4-metre Anglo-Australian Telescope at Siding Spring Observatory. The survey is made up of four volume-limited galaxy samples with the aim to cover a broad range in stellar mass and environment. The survey uses SAMI fibre 'hexabundles' to map these galaxies out to $>\sim 1$ effective radius. As the SAMI bundles have only 61 hexabundles to map the galaxies and with the chosen binning scale of 0.5" (see Sharp et al. 2015, for details), each SAMI galaxy has fewer spaxels than the S7 galaxies.

As with S7 we used eight galaxies in our training set, covering both strong and weak emission line galaxies and

1-comp	2-comp	3-comp
Total $\mathrm{Flux_{EM}}$/$\mathrm{Flux_{EM}}$_err	Total $\mathrm{Flux_{EM}}$/$\mathrm{Flux_{EM}}$_err	Total$\mathrm{Flux_{EM}}$/$\mathrm{Flux_{EM}}$_err
	Comp1 $\mathrm{Flux_{EM}}$/Comp1 $\mathrm{Flux_{EM}}$_err	Comp1 $\mathrm{Flux_{EM}}$/Comp1 $\mathrm{Flux_{EM}}$_err
	Comp2 $\mathrm{Flux_{EM}}$/Comp2 $\mathrm{Flux_{EM}}$_err	Comp2 $\mathrm{Flux_{EM}}$/Comp2 $\mathrm{Flux_{EM}}$_err
		Comp3 $\mathrm{Flux_{EM}}$/Comp3 $\mathrm{Flux_{EM}}$_err
	Comp1 $\mathrm{Flux_{EM}}$/Total $\mathrm{Flux_{EM}}$	Comp1 $\mathrm{Flux_{EM}}$/Total $\mathrm{Flux_{EM}}$
	Comp2 $\mathrm{Flux_{EM}}$/Total $\mathrm{Flux_{EM}}$	Comp2 $\mathrm{Flux_{EM}}$/Total $\mathrm{Flux_{EM}}$
		Comp3 $\mathrm{Flux_{EM}}$/Total $\mathrm{Flux_{EM}}$
Comp1 $\sigma/\sigma_{\mathrm{err}}$	Comp1 $\sigma/\sigma_{\mathrm{err}}$	Comp1 $\sigma/\sigma_{\mathrm{err}}$
	Comp2 $\sigma/\sigma_{\mathrm{err}}$	Comp2 $\sigma/\sigma_{\mathrm{err}}$
		Comp3 $\sigma/\sigma_{\mathrm{err}}$
χ^2/DOF	χ^2/DOF	χ^2/DOF

Table 1. This table presents each input parameter given to The Machine as the input vector. Each parameter subscripted with EM are calculated for each of the following emission lines: Hα, Hβ, [*N*II]λ6583, [*S*II]$\lambda\lambda$6716, 6731, [*O*III]λ5007.

Seyfert and star-forming galaxies. Each spaxel in these eight galaxies were manually classified by five astronomers in the same manner as with S7. As the galaxies did not always fill the field of view of SAMI, pixels with no signal were left unclassified.

The SAMI fitting process includes the addition of errors in the continuum fit around Hα and Hβ to the Hα and Hβ flux errors. Outside of the galaxy, where there are no emission lines, this process removes the emission line fluxes.

From the five trainers, we label a spaxel by calculating the most common classification between the five astronomers. This, again, gives us a very clean sample of $\sim$ 2500 spaxels to train and test The Machine with. Again we don't find a 100% agreement between trainers but a $\sim$ 50% agreement between trainers. The increased number of trainers corresponds to the lower percentage of agreement between them, than we see with S7.

6 ACCURACY

We assessed the accuracy of The Machine after training using the subset of trainer classified input examples set aside for testing and cross-validation. We define the accuracy (how well The Machine matches the labels to our Trainers) of the machine as its ability to recall the same classifications as our trainers and the precision in making its decisions.

Equations 5 and 6 show how the recall (R) and precision (P) values are calculated for each number of components. N is the number of examples of which M (The Machine) and T (our trainers) classify with the conditions for M and T. Together these describe how well The Machine can classify examples in comparison to our trainers. These values are calculated for each component classification,

$$\mathrm{R_M} = \frac{\mathrm{N_{M=T}}}{\sum \mathrm{N_{M,T=1,2,3}}} \tag{5}$$

$$\mathrm{P_M} = \frac{\mathrm{N_{M=T}}}{\sum \mathrm{N_{M=1,2,3,T}}} \tag{6}$$

More completely, the recall $\mathrm{R_M}$ where M refers to The Machine, measures the consistency The Machine has for each classification related to how often it misclassifies an example of that component number. For example, if The Machine correctly classifies 200 examples as 1 components but misclassifies 50 1 component examples as 2 or 3 components, it has a recall of $\mathrm{R_1} = 200/250 = 75\%$ for 1 component classifications. A recall value is calculated for each classification, 1 to 3 for SAMI and 0 to 3 for S7. The precision $\mathrm{P_M}$ measures how often The Machine will misclassify an example as a particular number of components. For example, if The Machine correctly classifies 200 examples as 1 components but also incorrectly classifies 25 examples (of 2 and 3 components) as 1 component then The Machine has a precision of $\mathrm{P_1} = 200/225 = 89\%$ for 1 component classifications. Precision values are, like recall values, calculated for each classification.

This same comparison can be repeated using our individual trainers to show how The Machine's performance compares to astronomers visual inspections. Taking the training set of galaxies, we formed new component maps from N-1 trainers' classifications. For SAMI, this means we created 5 new combined classification maps using 4 of the 5 trainers for each one successively. For S7, we created new combined maps using 2 of the 3 trainers successively. We see a large spread in the agreement of classifications through the recall and precision values, as is shown in figure 4 by the solid lines. The dashed lines in figure 4 present The Machine's recall and precision values for each of the two surveys. We see that the largest spread in the ability of people agreeing with each other is between 2 and 3 components, while agreement is very good over 1 component fits. We have also shown that The Machine does as well at making the classification decisions for spectra as using individual people.

The Machine is able to define differences between classifications based on our training sets. To show this explicitly figures 5 and 6 present the component maps defined by The Machine and the trainers for a S7 galaxy and a SAMI galaxy, respectively. These galaxies were not used for training or cross-validation. We see that The Machine defines a component map which is between all of our trainers maps in both the SAMI and S7 galaxies.

The Machine's recall and precision is as good as our human trainers. On the full training sets for SAMI and S7, each trainer selected classifications based on their own biases to what they are seeing. By using those spaxels for which all trainers agreed we avoid these individual biases. Figure 7 shows the number of spaxels that each trainer classifies as each number of components for the full training and testing set for SAMI and S7. The Machine classifications are also shown to demonstrate the biases of the ANN. In both cases

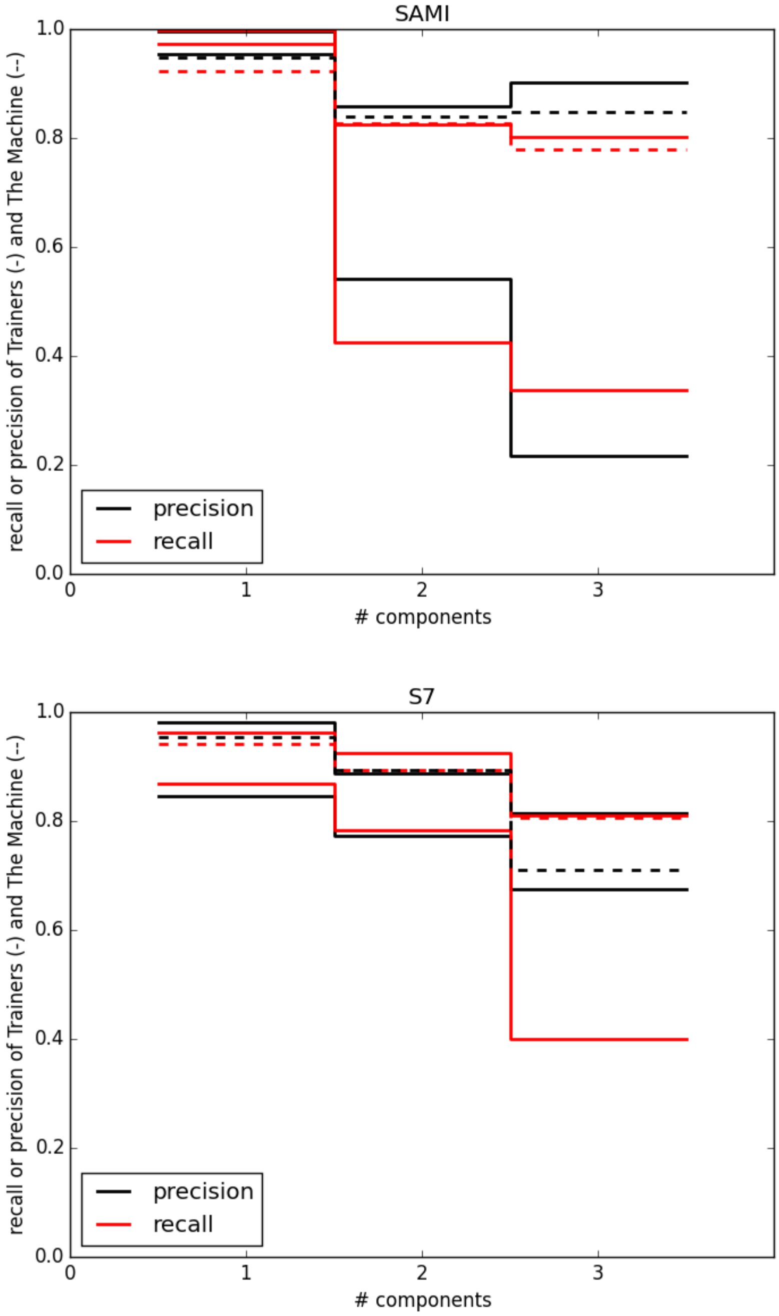

Figure 4. Here we present the minimum and maximum values of the precision and recall of our sets of trainers to the precision and recall of The Machine for SAMI (top) and S7 (bottom). The solid black lines show the min and max values of precision for each component number, the solid red lines indicate the min and max values of the recall for each component number. The dashed lines show The Machines results from training.

of SAMI and S7 we see that `The Machine` classifies the components in a similar manner to the trainers, following the average bias of the trainers as a whole. `The Machine` is biased towards 1 and 2 components but we can see that 2 of the 3 trainers are also biased towards 2 components over 1 components. This may also be the case that the S7 galaxies do have more 2 component spectra than 1 component spectra. S7 is selected to be very interesting Seyfert galaxies which we expect to require multi-component fits to express the data.

6.1 Comparison to an F-test

In addition to comparing `The Machine` to our trainers we have also compared the results of using an f-test to our trainers. Figure 8 presents the results of using an ftest on our training set of galaxies. The ftest selects 1-components

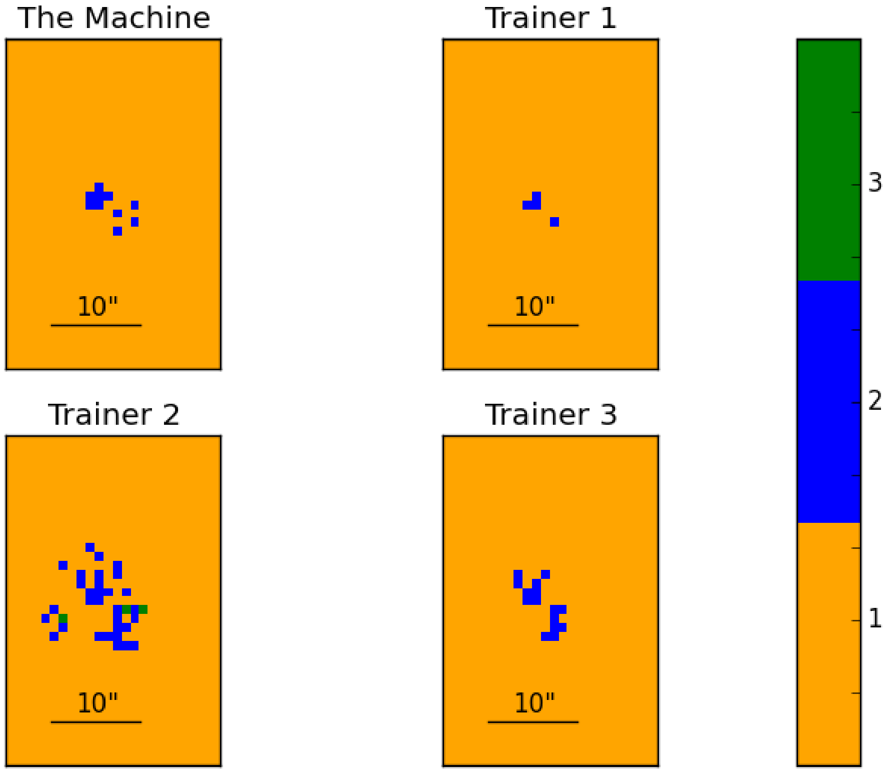

Figure 5. Component maps defined by The Machine and the three S7 trainers. Green for a 3-component fit, blue for a 2-component fit and yellow for a 1-component fit.

more often than our SAMI trainers. In addition, the precision obtained by using an ftest is comparable to people however the recall is much lower pointing to the fact that only ∼20% of the 1-component selections the ftest makes match to our trainers. Due to the discrepancies among our trainers on selecting 2, and 3 components the ftest is comparable in both recall and precision. In comparison to The Machine, however, the ftest is not as capable as representing another astronomer and their choices of classifications.

7 APPLICATION TO S7 AND SAMI

As presented in the previous sections, the LZIFU code (Ho et al. in prep), in combination with The Machine, can be used to provide a reliable decomposition of the different emission-line components present in galaxies observed with integral fields spectroscopy. With the availability of a reliable decomposition analysis for multiple emission-lines spanning the full optical spectrum and for each of our survey sources it becomes practical to undertake an in-depth analysis of the wide range of physical processes driving emission with complex composite sources. Early example of such analysis from the SAMI survey include phenomena such as binary black-holes (Allen et al. 2015b), metallicity measurements, corrected for underlying galaxy disk contamination, for isolated HII regions embedded within spiral galaxies (Richards et al. 2014), and the identification of shocks and outflows in modest luminosity star-forming galaxies (Ho et al. 2014, 2015).

Figure 9 shows the Hubble Space Telescope (HST) image of the Seyfert galaxy NGC7582, whose central regionhas been observed with WiFeS as part of S7 (blue rectangle in figure 9). This galaxy has a large star-forming disk, visible in the image. Perpendicular to this disk is an ionisation cone with an opening angle of 110 degrees that is excited by the central AGN, highlighted in figure 10 and described in Dopita et al. (2015). The gas within this cone is highly ionised and extends to 15kpc. The counter-cone is also visible in our optical observation but is partly obscured by the dust of the star forming disk. The red line on figure 9 indicates the major axis and the circle indicates the centre where the AGN is located. The S7 observation of NGC7582 has been fit with 1, 2, and 3 components using LZIFU and then run through The Machine, to obtain the merged component maps. The decomposition obtained with lzifu and The Machine is shown in figure 10 for NGC 7582.

The decomposition of emission lines into different components enables the separation of the different excitation processes occurring within a galaxy. In NGC7582, the decomposition of the emission lines separates the galactic disk from the ionisation cone and counter-cone. In figure 10, panels (a), (b), and (c) present the continuum map, 3-

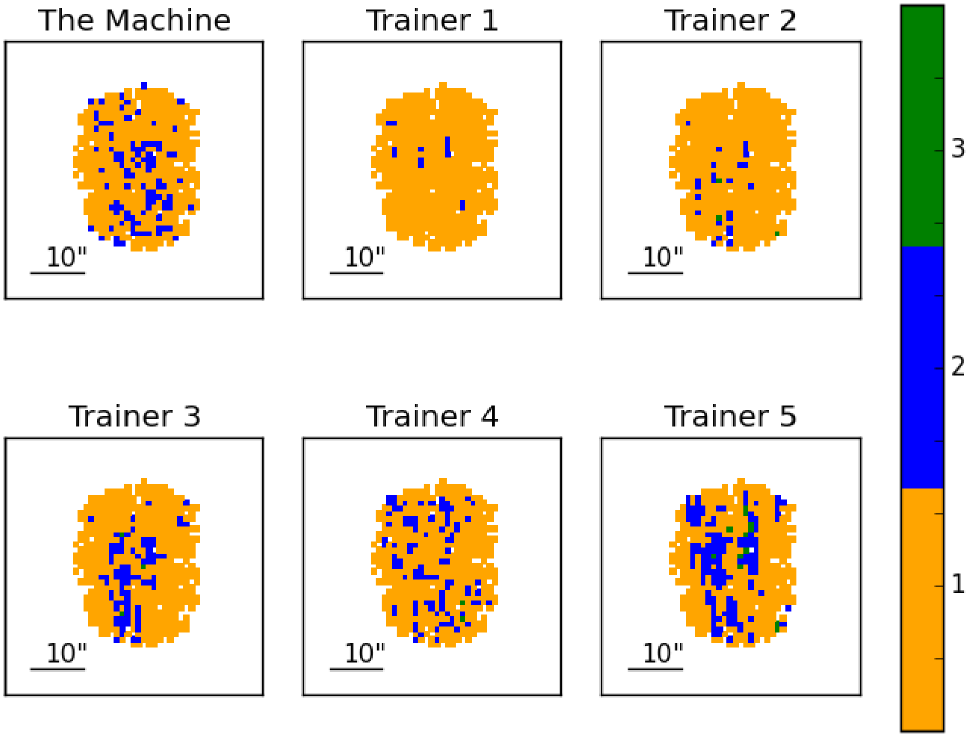

Figure 6. Component maps defined by `The Machine` and the five SAMI trainers. Green for a 3-component fit, blue for a 2-component fit and yellow for a 1-component fit.

colour emission-line flux map ([OIII], [NII] and Hα), and the [NII]/Hα total flux ratio map of NGC5782. Multi-component line fitting was necessary because these lines are not always well described by a single Gaussian. The decomposition of NGC7582 is presented in figure 10, panels (d), (e), and (f), which show the velocities assigned to each component for each spaxel. Component 1 contains the narrowest emission line components and traces the disk of the galaxy. We can see the rotation curve of this disk gas in panel (g) as we trace the axis of rotation from figure 9 along the galaxy. We do not see a turnover in the rotation curve because the S7 observations are looking at only the central regions of the galaxy. The second component shown in panel (e) consists of the broadest emission line components and traces the ionisation cone. We can verify this by looking at the velocity of the second component, as a function of distance from the centre, in the area of the cones. The counter-cone is partly obscured by the galactic disk; we see this in panel (h). The cone and counter-cone are both moving material at a projected velocity of +/- 100km/s. The velocity plateaus in panel (h) suggesting the front cone is outflowing. We see the mirror of this in what we suspect is the counter-cone, blue dashed lines. Further analysis is required but beyond the scope of this research. The remainder of the points in panel (h) are most likely due to the disk of the galaxy broadened due to beam smearing in our line of sight. The third component is a secondary narrow component of emission. In Figure 10 panel (g), the histogram of the velocity dispersions of each component, these third components are located between the first and second and are labelled in red. These spaxels have separated narrow peaks with a broader underlying component. These components may be due to the ionisation of matter around or at the edge of the cone. To determine what causes this third component, we have looked at the ionisation hardness of each component of each spaxel using the [NII] diagnostic diagram (Baldwin et al. 1981). Each component is plotted in a separate colour, this third component (red) shows the highest ionisation. This indicates that it may be shock-induced.

Although this paper does not go into further detail on NGC7582, we have shown that the decomposition of emission lines is important in understanding what is happening within a galaxy. McElroy et al. (2015) found it beneficial to fit each galaxy with `LZIFU` then use an f-test with harsh cut-offs to determine the component decompositions. In section 6.1, however, we have shown that the f-test does not select components like astronomers and may miss some complexities in some emission lines.

Surveys are now creating more data than before, making it not always feasible to fit emission lines by hand, nor to make the component decisions by eye. This is where `The Machine` is most advantageous. `The Machine` is able to pro-

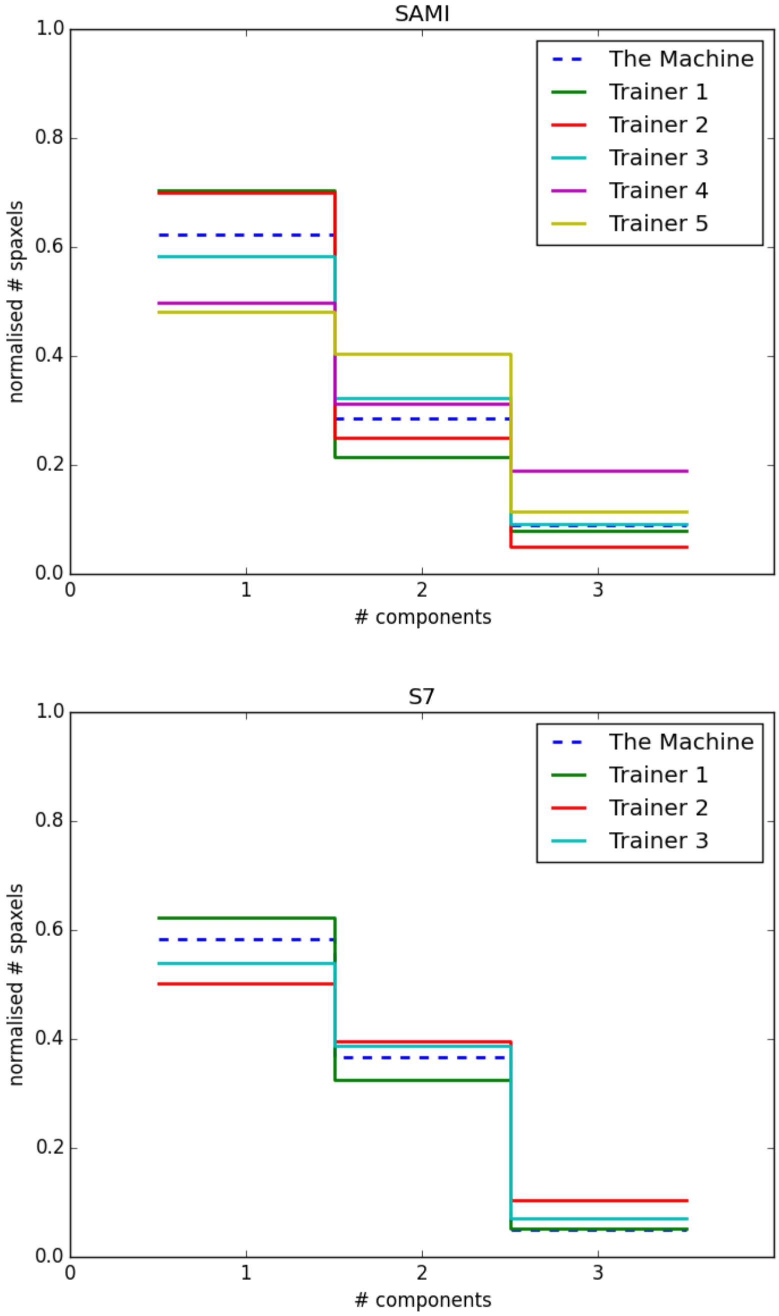

Figure 7. Top: A representation of how many 1, 2, or 3 component classifications made by each trainer (-) and The Machine (–) for the S7 training set of galaxies. Bottom: A representation of how many 1, 2, or 3 component classifications made by each trainer (-) and The Machine (–) for the SAMI training set of galaxies. In both training sets The Machine predicts similar numbers of 1, 2, or 3 components for the training sets.

cess thousands of galaxies and assign the best representation of emission line fits as well as an astronomer in very little time. This then allows the deeper analysis of galaxies such as NGC7582 through the multi-component emission line fits.

Figure 11 presents the results of running the SAMI Galaxy Survey DR1 data through The Machine. The left panel shows a histogram of the number of Gaussian components classified by The Machine for all spectra. This is a total of 348,023 spaxels, with the majority being one component fits. The right panel presents the the percentages of a galaxy that are described by 1, 2, or 3 Gaussian components. This allows us to pin-point galaxies that show mostly starformation (mostly 1-component fits) and those galaxies which present as having multiple physical processes ongoing (greater than zero percentage of 2 or 3 component fits).

In a further study, we will be looking at the prevalence of multicomponent emission lines in the SAMI Galaxy Survey. This will mean comparing the number of emission line

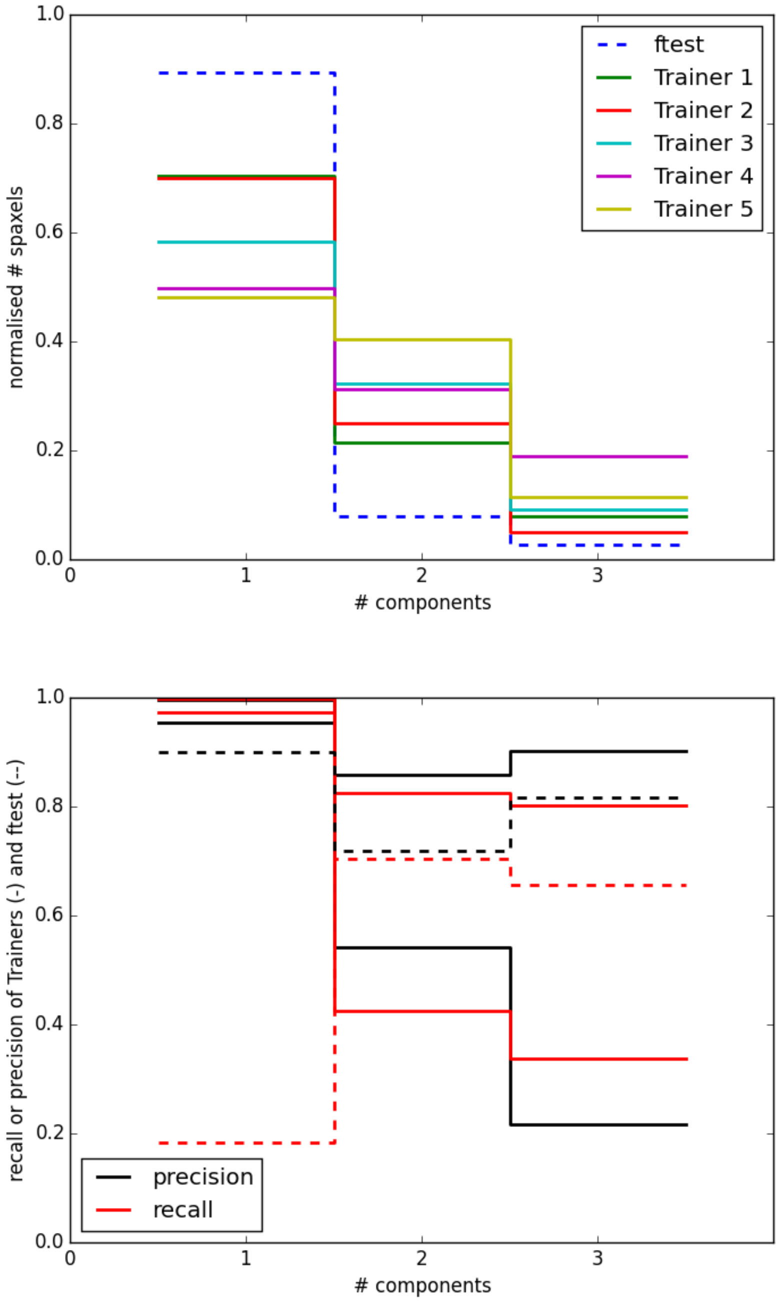

Figure 8. Top: A comparison of the number of 1, 2, and 3 components classified by the SAMI trainers (-) and using an ftest (–). Bottom: A comparison of the recall and precision obtained by using an ftest on the SAMI training set of galaxies in comparison to our trainers against each other.

components to the galaxy's mass, AGN activity, star formation history and other parameters, to look for correlations that may help in identifying certain types of galaxies or to help understand which types of galaxies contain combinations of multiple physical processes. A study of galaxy type and component fitting is only possible with hundreds or thousands of galaxies all fit with multiple components. Using an ANN to make the classifications of the emission line fits has made it possible to do this study on a short timescale. Instead of waiting years for an entire survey to be reduced to the point of multi-component emission line fits by hand, the reduction can be done in weeks, opening the possibility of statistical studies of multicomponent emission processes in a large range of galaxies.

8 CONCLUSION

Complex emission line fitting of spectra is not new. But with the larger IFU surveys now in progress, automated com-

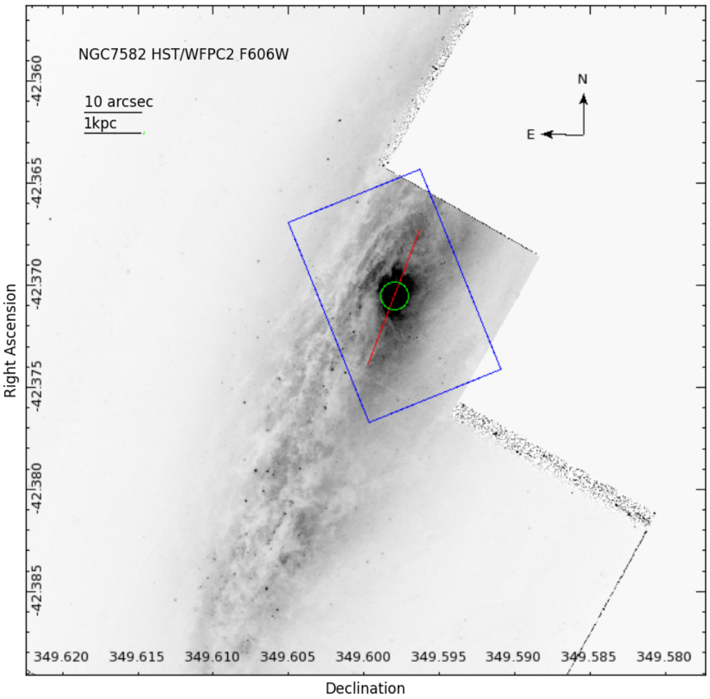

Figure 9. HST image of NGC5782. Blue box indicates the S7 FoV and the red line shows the major axis of the galaxy we have used in our analysis of the WiFeS data. The green circle indicates the centre of the galaxy.

plex emission line fitting is a must. `LZIFU` has automated the fitting process for up to 3 Gaussian components, but does not have the capacity to determine how many components are the best for a particular spaxel. Our ANN, `The Machine`, can, indicating that the complexities of differentiating between multi-component fits can be solved reliably and rapidly.

We have built `The Machine` to take in information produced by LZIFU and output the best fit classification to each individual spaxel in each galaxy of a survey. It is a fast, self-consistent, and reliable system that replaces the need for years of manual work by astronomers. The breakdown of the accuracy into recall and precision of `The Machine` shows that it is indistinguishable from our human trainers.

Our analysis shows that an ANN trained by sets of astronomers is capable of classifying new galaxies to the same reliability as another astronomer. The only difference is that is does not need sleep, food, a break, or to be paid. `The Machine` is able to consistently, reliably, and quickly classify spaxels as needing 1, 2, or 3 Gaussian components so that astronomers can focus on the analysis of the emission line fluxes, velocity dispersions, and velocities to determine what is going on in the galaxies in their surveys.

ACKNOWLEDGMENTS

REFERENCES

Allen J. T. et al., 2015a, MNRAS, 446, 1567
Allen J. T. et al., 2015b, MNRAS, 451, 2780
Baldwin J. A., Phillips M. M., Terlevich R., 1981, PASP, 93, 5
Bobra M. G., Couvidat S., 2015, ApJ, 798, 135
Bundy K. et al., 2015, ApJ, 798, 7
Croom S. M. et al., 2012, MNRAS, 421, 872
Doert M., Errando M., 2014, ApJ, 782, 41
Dopita M. et al., 2010, APSS, 327, 245
Dopita M. A. et al., 2014, AAP, 566, A41
Dopita M. A. et al., 2015, ApJS, 217, 12
Driver S. P. et al., 2009, Astronomy and Geophysics, 50, 12
Hampton E., et al., in prep
Hassan T. et al., 2013, MNRAS, 428, 220
Ho I.-T. et al., 2014, MNRAS, 444, 3894
Ho I.-T. et al., 2015, MNRAS, 448, 2030
Ho I.-T., et al., in prep
Husemann B. et al., 2013, A&A, 549, A87
Kuminski E. et al., 2014, PASP, 126, 959
McElroy R. et al., 2015, MNRAS, 446, 2186
Richards S. N. et al., 2014, MNRAS, 445, 1104
Sánchez S. F. et al., 2012, A&A, 538, A8
Sharp R. et al., 2015, MNRAS, 446, 1551

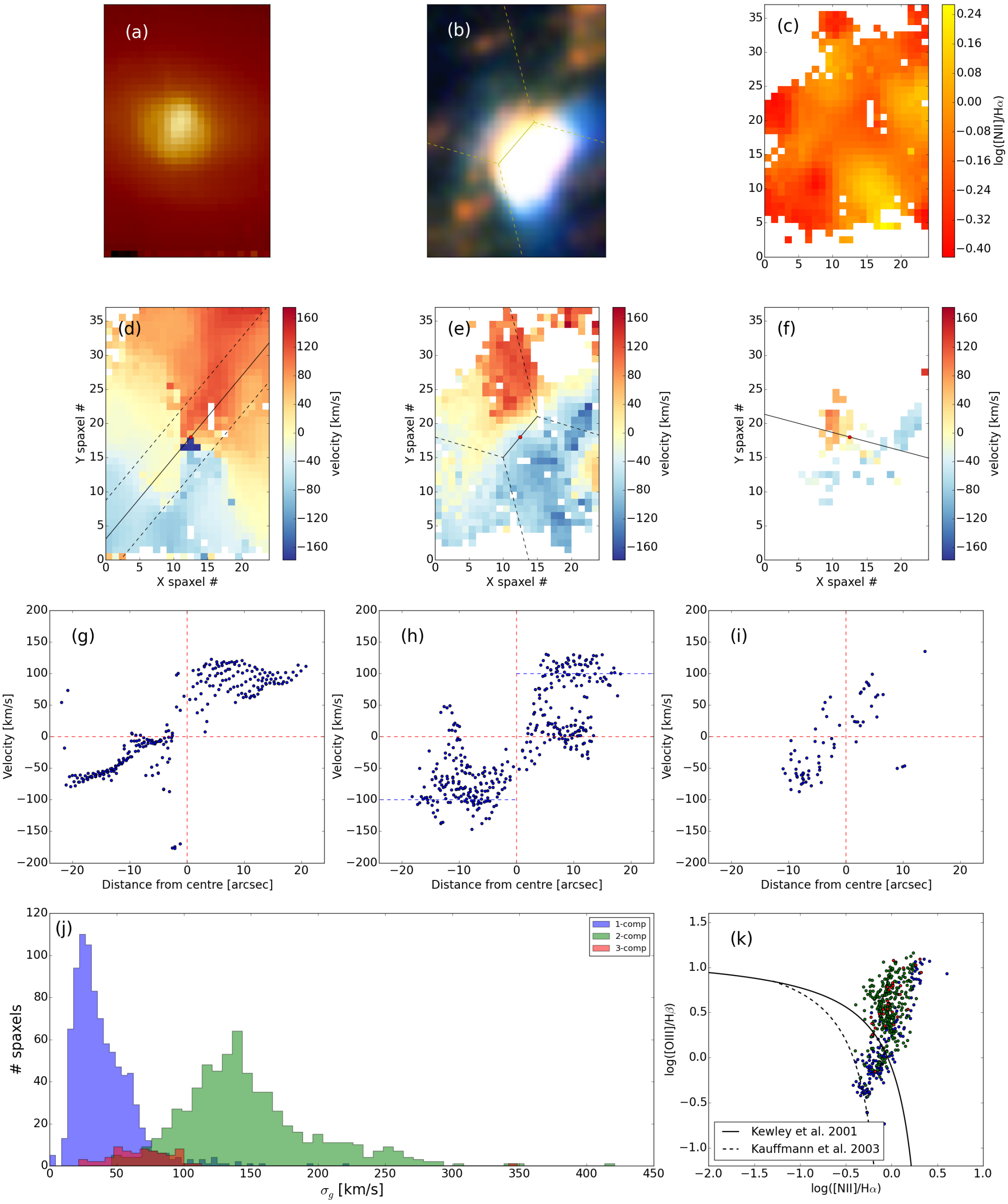

Figure 10. (a) Continuum image (red) of NGC7582. (b) [OIII] - blue, [NII] - green, Hα - red. (c) Total [NII]/Hα flux map. (d) Velocity field of the 1st component related to the disk. (e) Velocity field of the 2nd component related to the ionisation cone and counter-cone. (f) velocity field of the 3rd component related to the interaction at the edge of the ionisation cone. (g) Rotation curve of the disk gas, points taken from 1st component velocity within dashed lines of (d). (h) Rotation of gas due to the ionisation cones. Plateau at +/- 100 km/s indicative of the outflowing gas. Contamination from dust obscured emission. (i) rotation of 3rd component. (j) Velocity dispersions of each component colour-coded by which component. (k) [NII] BPT diagram with components colour-coded to show that the 3rd component has high ionisation as is expected from being caused by the interaction of the ionisation cone.

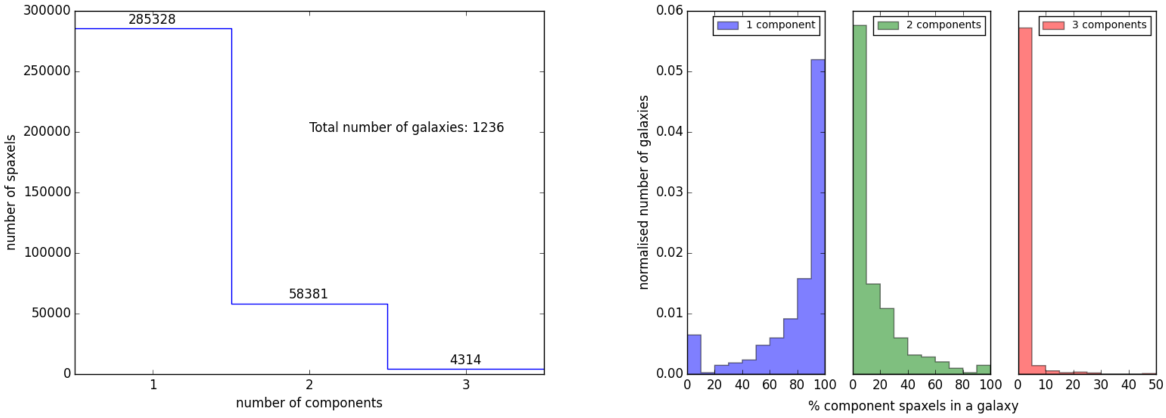

Figure 11. The left histogram presents the number of components classified for the 1236 current SAMI galaxies. The right presents the numbers of galaxies with different percentages of 1, 2 or 3 component spaxels. While a significant fraction of SAMI galaxies reveal a 2nd component in many of their spaxels, almost no galaxies (%) have any spaxels with a 3rd component.