The Astrophysical Journal, 883:63 (13pp), 2019 September 20 https://doi.org/10.3847/1538-4357/ab397e
© 2019. The American Astronomical Society. All rights reserved.

Machine-learning Classifiers for Intermediate Redshift Emission-line Galaxies

Kai Zhang[1], David J. Schlegel[1], Brett H. Andrews[2], Johan Comparat[3,4,5], Christoph Schäfer[6], Jose Antonio Vazquez Mata[7], Jean-Paul Kneib[6,8], and Renbin Yan[9]
[1] Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA; zkdtckk@gmail.com
[2] PITT PACC, Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA 15260, USA
[3] Instituto de Física Teórica UAM/CSIC, E-28049 Madrid, Spain
[4] Departamento de Física Teórica, Universidad Autónoma de Madrid, E-28049 Madrid, Spain
[5] Max-Planck-Institut für extraterrestrische Physik (MPE), Giessenbachstrasse 1, D-85748 Garching bei München, Germany
[6] Institute of Physics, Laboratory of Astrophysics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland
[7] Instituto de Astronomía, Universidad Nacional Autónoma de México, A.P. 70-264, 04510, Mexico, D.F., México
[8] Aix Marseille Université, CNRS, LAM (Laboratoire d'Astrophysique de Marseille) UMR 7326, F-13388, Marseille, France
[9] Department of Physics and Astronomy, University of Kentucky, 505 Rose Street, Lexington, KY 40506, USA
Received 2019 April 24; revised 2019 July 17; accepted 2019 August 6; published 2019 September 20

Abstract

Classification of intermediate redshift ($z = 0.3$–0.8) emission line galaxies as star-forming galaxies, composite galaxies, active galactic nuclei (AGNs), or low-ionization nuclear emission regions (LINERs) using optical spectra alone was impossible because the lines used for standard optical diagnostic diagrams: [N II], Hα, and [S II] are redshifted out of the observed wavelength range. In this work, we address this problem using four supervised machine-learning classification algorithms: *k*-nearest neighbors (KNN), support vector classifier (SVC), random forest (RF), and a multilayer perceptron (MLP) neural network. For input features, we use properties that can be measured from optical galaxy spectra out to $z < 0.8$—[O III]/Hβ, [O II]/Hβ, [O III] line width, and stellar velocity dispersion—and four colors ($u - g$, $g - r$, $r - i$, and $i - z$) corrected to $z = 0.1$. The labels for the low redshift emission line galaxy training set are determined using standard optical diagnostic diagrams. RF has the best area under curve score for classifying all four galaxy types, meaning the highest distinguishing power. Both the AUC scores and accuracies of the other algorithms are ordered as MLP > SVC > KNN. The classification accuracies with all eight features (and the four spectroscopically determined features only) are 93.4% (92.3%) for star-forming galaxies, 69.4% (63.7%) for composite galaxies, 71.8% (67.3%) for AGNs, and 65.7% (60.8%) for LINERs. The stacked spectrum of galaxies of the same type as determined by optical diagnostic diagrams at low redshift and RF at intermediate redshift are broadly consistent. Our publicly available code (https://github.com/zkdtc/MLC_ELGs) and trained models will be instrumental for classifying emission line galaxies in upcoming wide-field spectroscopic surveys.

Key words: galaxies: active – galaxies: Seyfert – quasars: emission lines

1. Introduction

Accurate classification of emission line galaxies is critical because the different types of emission line galaxies correspond to different underlying excitation and ionization conditions. Applying an analysis technique intended for one type of galaxy on another type can produce qualitatively incorrect results (e.g., applying a metallicity calibration on an active galactic nucleus; AGN) because of the built-in assumptions about excitation and ionization conditions.

Standard optical diagnostic diagrams, such as the BPT (Baldwin et al. 1981) or VO87 (Veilleux & Osterbrock 1987) diagrams, are widely used to classify low redshift emission line galaxies into star-forming galaxies (SFGs), composite galaxies, AGNs, and LINERs. These diagnostic diagrams use the [O III]/Hβ, [N II]/Hα, [S II]/Hα, and/or [O I]/Hα lines ratios and some demarcation criteria (e.g., Kauffmann et al. 2003; Kewley et al. 2006). The advent of large optical spectroscopic surveys like the Sloan Digital Sky Survey (SDSS; York et al. 2000), 2dF (Boyle et al. 2000), and LAMOST has enabled the classification of hundreds of thousands of low redshift ($z < 0.3$) emission line galaxies.

Classifying intermediate ($z > 0.3$) emission line galaxies is significantly more difficult because the optical spectral features used in the BPT diagram are not captured in optical spectra at these redshifts. Obtaining the rest-frame optical spectra to apply the BPT diagram requires getting rare and expensive infrared spectra (Kewley et al. 2013a, 2013b; Trump et al. 2013; Azadi et al. 2017). Classifying intermediate redshift galaxies using only optical spectral and photometric information will enable a wide range of emission line galaxy science with upcoming Stage-IV optical spectroscopic surveys like Dark Energy Spectroscopic Instrument (DESI, Levi et al. 2013), Subaru Prime Focus Spectrograph (PFS; Takada et al. 2014; Tamura et al. 2016), and the 4 m Multi-Object Spectroscopic Telescope (4MOST; de Jong et al. 2012).

Currently, there are dozens of classification diagrams developed only using parameters available from optical spectra. Typically, these methods use the fact that AGNs reside exclusively in massive, fast-rotating galaxies and have strong high-ionization lines while SFGs are less massive, rotate slower, and have lower ionization states. Some examples of these diagrams include:

1. the EW([O II]) versus EW([O III]) diagram (Tresse et al. 1996; Rola et al. 1997);
2. the DEW diagram, which uses 4000 Å break $D_n(4000)$, EW([O II]λ3727) and EW([Ne III]λ3870) (Stasińska et al. 2006);
3. diagrams using $g - z$, [Ne III], and [O II] (Trouille et al. 2011);

4. the H-band absolute magnitude versus [O III]/Hβ diagram (Weiner et al. 2006);
5. the [O II]/Hβ versus [O III]/Hβ diagram (Lamareille 2010);
6. the $U-B$ color versus [O III]/Hβ diagram (Yan et al. 2011);
7. the mass–excitation diagnostic (MEx), which uses the stellar mass versus [O III]/Hβ diagram (Juneau et al. 2011, 2013);
8. the $D_n(4000)$ versus [O III]/Hβ diagram (Marocco et al. 2011); and
9. the kinematic–excitation diagram (KEx), which uses [O III] line width versus [O III]/Hβ (Zhang & Hao 2018).

These diagnostic diagrams generally separate SFGs and AGNs well, but none of them classify emission line galaxies into the four subtypes that the BPT produces: SFGs, composite galaxies, AGNs, and LINERs. Composite galaxies and LINERs are heavily mixed with SFGs or AGNs on these diagrams. In this work, we explore the potential for machine-learning algorithms to provide accurate four-class classifications using input features from optical spectra and photometric colors.

In recent years there has been an explosion in the number of applications of machine-learning techniques to astronomical problems (see Acquaviva 2019 for a review). While some studies have used unsupervised algorithms (e.g., Hocking et al. 2018), supervised algorithms have proven to be even more powerful. The accuracy of neural networks, especially deep convolutional neural networks, to classify astronomical images has improved dramatically since the early work by de la Calleja & Fuentes (2004). For instance, Dieleman et al. (2015) used a deep convolutional neural network for classifying galaxies using human-labeled images from the Galaxy Zoo project (Lintott et al. 2011) that outperforms experts. Deep neural networks are also well-suited for identifying strong lens systems in galaxy images because these systems can be robustly simulated even though they are rare in nature (Jacobs et al. 2017, 2019a, 2019b; Petrillo et al. 2017; Metcalf et al. 2018; Pourrahmani et al. 2018; Huang et al. 2019).

Despite the excitement surrounding deep convolutional neural networks, classical supervised machine-learning algorithms are often more accurate for problems with relatively few input features, such as classifying emission line galaxies from optical spectral features and photometric colors. In this paper, we use several such algorithms: K-nearest neighbors (KNN), support vector classifier (SVC), random forest (RF), and a multilayer perceptron neural network (MLP-NN).

The layout of the paper is as follows. Section 2 describes the selection and labeling of training, test, and target samples. Section 3 discusses the selection of input features. Section 4 compares the performance of our four supervised learning algorithms for classifying low redshift emission line galaxies. Section 5 describes the application of the trained models to intermediate redshift galaxies. Section 6 contains our main conclusions. We use a cosmology with $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, $\Omega_\mathrm{m} = 0.3$, and $\Omega_\Lambda = 0.7$ throughout this paper.

2. Sample

2.1. z < 0.32 Training and Test Samples

We apply the following criteria to the SDSS-IV DR15 (Blanton et al. 2017; Aguado et al. 2019) Extended Baryon Oscillation Spectroscopic Survey (eBOSS; Dawson et al. 2016) data for selecting the low redshift sample for model training, validation, and testing. The Value Added Catalog[10] is used to get all the spectral and photometric data used here. We use the Python implementation[11] of the kcorrect package (Blanton & Roweis 2007) to convert the *u, g, r, i, z* magnitudes to $z = 0.1$ values. In order to get the appropriate SED for a set of galaxy fluxes, kcorrect fits an SED, which is a nonnegative linear combination of some small number of carefully chosen templates. The $z < 0.32$ emission line galaxies sample is selected according to the following criteria:

$0 < z < 0.32$
CLASS = "GALAXY"
SN([O II]) > 3, SN(Hβ) > 3, SN([O III]) > 3, SN([N II] λ6583) > 3, SN(Hα) > 3, SN([S II] $\lambda\lambda$6717, 6731) > 3
[O III] emission line width (σ([O III])) > 0, stellar velocity dispersion (σ_*) > 0.

The final sample consists of 28,869 galaxies, which we split 70/30 into a training+validation sample (20,208 galaxies) and a test sample (8661 galaxies). We use k-fold splitting ($k = 6$ for this work) to divide the training+validation sample, which means the sample is split into six equal subsamples and each time one subsample is used as validation sample while the remaining five are training samples. This method gives an estimation of error introduced by sample variation and reduces this error in the final prediction by averaging over all k models.

2.2. Data Labels

At $z < 0.32$, galaxies can be classified into SFGs, composite galaxies, AGNs, and LINERs using BPT diagrams (Baldwin et al. 1981; Veilleux & Osterbrock 1987; Kauffmann et al. 2003; Kewley et al. 2006). We use the demarcation lines proposed in Kauffmann et al. (2003) and Kewley et al. (2006) for classification into four subtypes. The distributions of the four subtypes of galaxies in the BPT diagrams are shown in Figure 1. SFGs, composites, AGNs, and LINERs are denoted in blue, green, red, and orange, respectively. There are 17,073 SFGs, 6826 composites, 2566 AGNs, and 2399 LINERs.

2.3. 0.32 < z < 0.8 Emission Line Galaxies Sample

Our goal is to classify intermediate redshift ($0.32 < z < 0.8$) ELGs. The intermediate redshift sample is selected based on the following criteria:

$0.32 < z < 0.8$
CLASS = "GALAXY"
SN([O II]) > 3, SN(Hβ) > 3, SN([O III]) > 3
[O III] emission line width (σ([O III])) > 0, stellar velocity dispersion (σ_*) > 0.

The final sample consists of 49,272 galaxies. We use the kcorrect package to convert the *u, g, r, i, z* magnitudes to $z = 0.1$ values.

3. Input Features

In machine-learning terminology, features are the input parameters. We use "features" as the standard term here. We select [O III]/Hβ, [O II]/Hβ, [O III] λ5007 line width $\sigma_{[\mathrm{O\,III}]}$,

[10] https://data.sdss.org/sas/dr14/eboss/spectro/redux/v5_10_0/
[11] https://pypi.org/project/kcorrect_python/

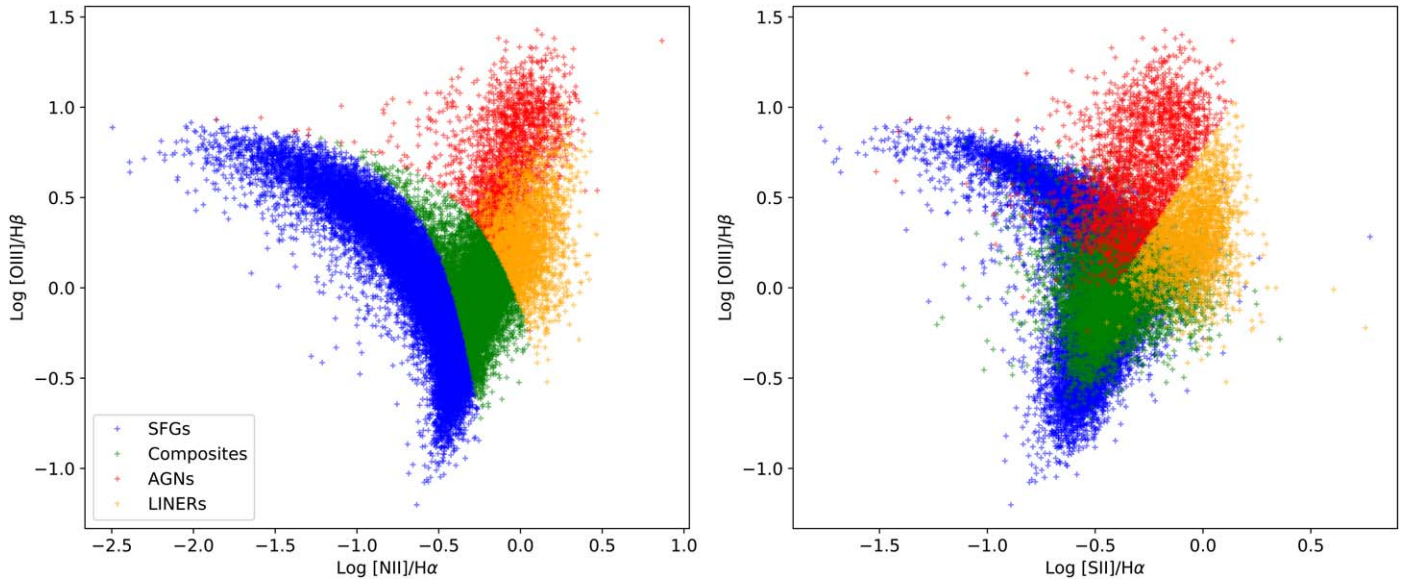

Figure 1. Classification of star-forming galaxies (blue), composites (green), AGNs (red), and LINERs (orange) using the BPT diagram for the $z < 0.32$ galaxy sample. The demarcation lines are from Kauffmann et al. (2003) and Kewley et al. (2006).

stellar velocity dispersion σ_*, $u - g$, $g - r$, $r - i$, and $i - z$ as the input features for classification. [O III]/Hβ, [O II]/Hβ, [O III] λ5007 line width $\sigma_{[\mathrm{O\ III}]}$, stellar velocity dispersion σ_* can be easily measured from optical spectra of $z < 0.8$ galaxies and can be measured out to even higher redshift if NIR spectra are available. The SDSS imaging survey[12] provides $u - g$, $g - r$, $r - i$, and $i - z$ colors for 14,055 square degrees of the sky. If a source is not detected, we use its magnitude upper limit because an upper limit is still informative. The g, r, and z photometry is supplemented using The Legacy Surveys[13] Data Release 7 (Dey et al. 2019) values if available. The Legacy Surveys are producing an inference model catalog of the sky from a set of optical and infrared imaging data, comprising 14,000 deg^2 of the extragalactic sky visible from the northern hemisphere in three optical bands (g, r, and z) and four infrared bands. These input features are selected from previous works of intermediate redshift emission line galaxies diagnostic diagrams (e.g., Lamareille 2010; Yan et al. 2011; Zhang & Hao 2018). They are chosen because [O II], Hβ, and [O III] are the strongest emission lines at rest-frame wavelengths shorter than 5010 Å. Stellar velocity dispersion can be well-measured using continuum fitting, and the u, g, r, i, and z broadband magnitudes have high signal-to-noise ratios. We do not use stellar mass measurements (Juneau et al. 2011) because these are derived values with typical errors of 0.3–0.4 dex, and σ_* and $\sigma_{[\mathrm{O\ III}]}$ already contain information about the mass of a galaxy. We chose not to use $D_n(4000)$ because it is less informative than the $u - g$ color. The [Ne III] line is not selected because of its weakness. One could add more input features, like colors using other bands, more emission line ratios, or equivalent widths, and this might or might not improve the classification accuracy. For this paper, we just use the eight input features to set a baseline.

In Figure 2, we show the distribution of the eight features for the four subtypes for the low redshift galaxy sample. Figure 3 shows the median values of the eight input features for the four subtypes ELGs for the whole $z < 0.32$ sample to illustrate the distinguishing power of each feature. All features are normalized to the 5–95 percentile range. SFGs are characterized by low [O III]/Hβ, [O II]/Hβ, $\sigma_{[\mathrm{O\ III}]}$, σ_*, $u - g$, $g - r$, and $r - i$. Thus, they are clustered in a very small volume in the 8D parameter space. AGNs are characterized by extremely high [O III]/Hβ, and all other seven features are near the median. LINERs show the highest [O II]/Hβ, σ_*, and $g - r$ color. Composites have median values for all eight features between 0.4 and 0.6. On average, the four subtypes are easily distinguished using the eight features. However, we do not consider the dispersion of each feature, so the four subtypes could still be heavily mixed with each other in parameter space and thus not 100% separable, as shown later in the paper.

4. Model Training and Performance

We use several popular supervised learning methods and quantify their classification accuracy. For each method, we briefly introduce the algorithm, fine-tune the hyperparameters, and report its performance. We note that performance on our data set is not necessarily indicative of performance on other data sets because these methods are sensitive to the particulars of the data set. The discussion is strictly confined to the data we use here and the models we use.

The four subtypes of emission line galaxies are not equally represented in the final sample. There are 17,073 SFGs, 6826 composites, 2566 AGNs, and 2399 LINERs. If we directly feed the imbalanced training sample into a model, it will be biased in favor of the over-represented subtypes and biased against the under-represented subtypes. In our case, the model would excel at selecting SFGs but struggle with distinguishing the other three subtypes. To mitigate this problem, we created a new sample equally weighted across subtypes by randomly selecting galaxies from each subtype. This is equivalent to giving higher weights to subtypes with fewer instances.

To quantify the performance of each method, the trained classifier is applied to the test sample, and the fraction of correct classification is the accuracy for a specific subtype. The receiver operating characteristic (ROC) curve (Metz 1978;

[12] https://www.sdss.org/dr12/imaging/
[13] http://legacysurvey.org

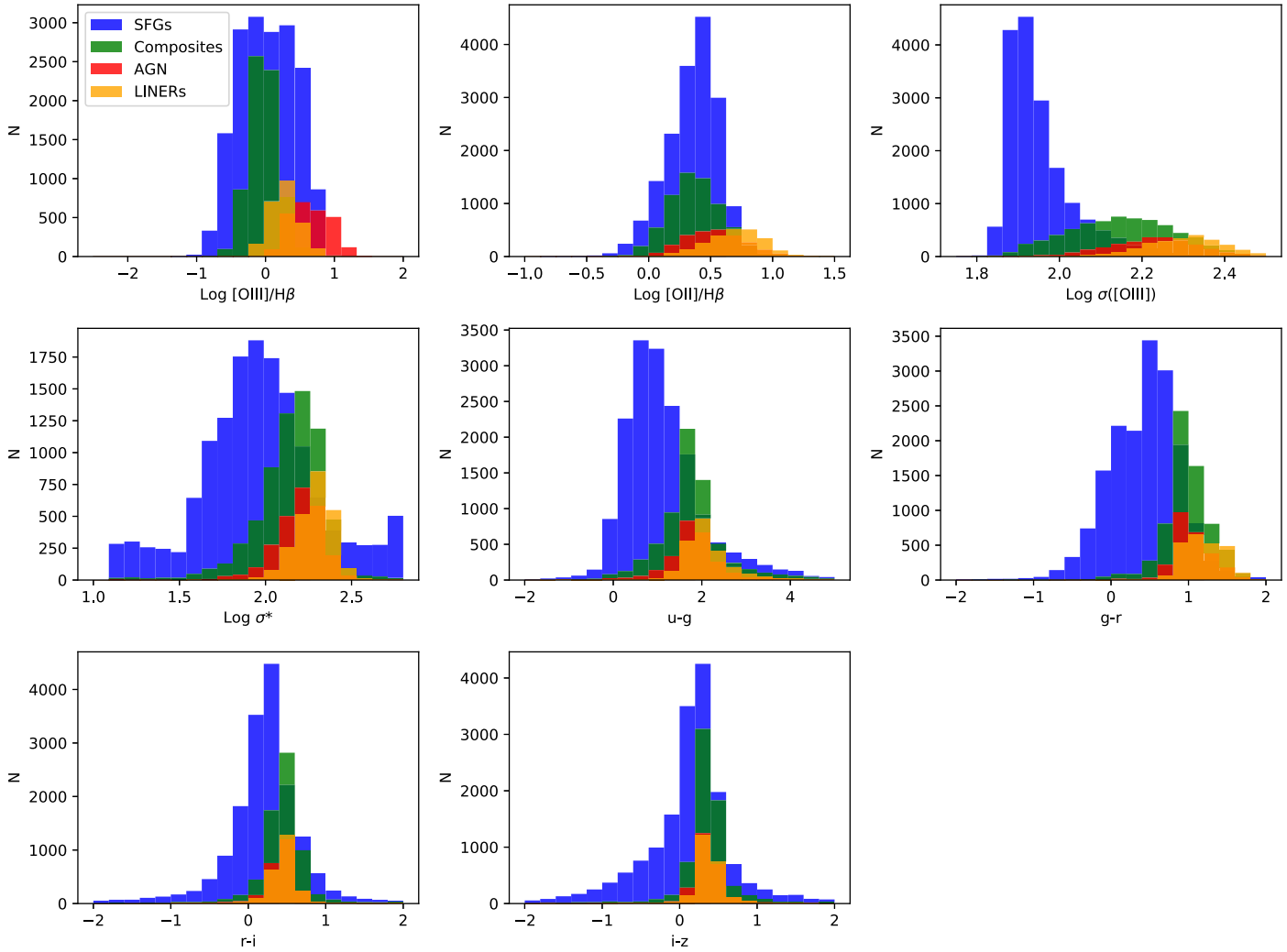

Figure 2. Distribution of the eight input features for the four subtypes of ELGs for the whole $z < 0.32$ sample.

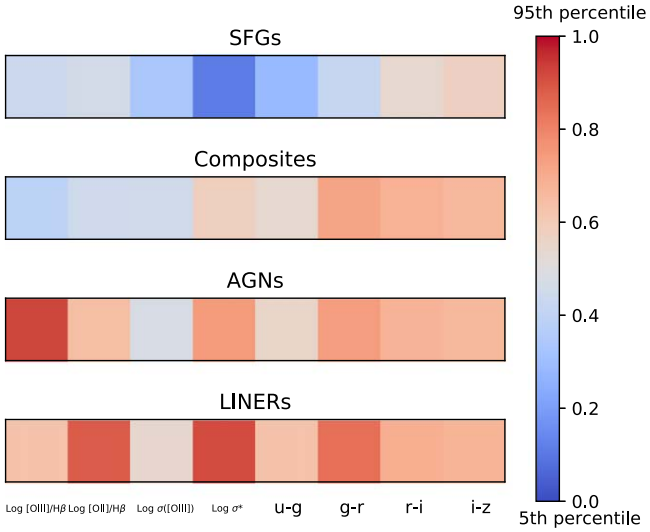

Figure 3. Median values of the eight input features for the four subtypes of ELGs for the whole $z < 0.32$ sample to illustrate the distinguishing power of each feature. All features are normalized to the 5–95 percentile range. The median values of the four subtypes are easily distinguished from each other using the eight features here.

Fawcett 2006) and area under the ROC curve (AUC) score (Bradley 1997) are used to quantify the distinguishing power of different classifiers. The confusion matrix of a classifier include:

True Positive (TP)—correct identification.
True Negative (TN)—correct rejection.
False Positive (FP)—incorrect identification, also called a false alarm or Type I error.
False Negative (FN)—incorrect rejection, also called a Type II error.

The ROC curve uses the true positive rate $\left(\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}\right)$ as the y-axis and the false positive rate $\left(\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}\right)$ as the x-axis. It reflects the tradeoff between TPR and FPR for different thresholds and thus different demarcation hyperplanes. In binary classification, setting the threshold too high will produce a high TPR and high FPR, meaning that a large fraction of true positives get selected but a large number of false positives will be misclassified as well. Setting the threshold too low results in a low TPR and a low FPR, meaning that many false positives are successfully rejected at the cost of not selecting many true positives. A perfect classifier has a TPR = 1 and an FPR = 0, which produces an AUC score of 1. The worst possible classifier, on the other hand, has a TPR = 0 and an FPR = 1,

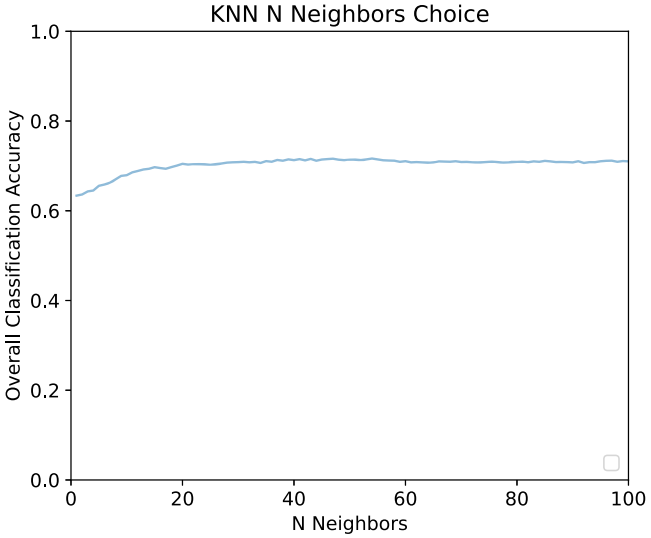

Figure 4. Overall accuracy on the validation sample as a function of the number of neighbors for classification voting (k). The best performance is achieved with $k = 54$. Therefore, k is set to 54.

which produces an AUC score of 0. Consequently, AUC score is commonly used to evaluate the classification power of a model with higher AUC scores being better. As such, this tool has become standard in optimization scenarios, but applying it to multiclass cases is more challenging. The general idea is to convert the multiclass problem into several binary classification problems using the one-versus-rest method (Mossman 1999; Srinivasan 1999; Hand & Till 2001; Ferri et al. 2009). For each ML technique we describe in the following sections, we present the ROC curves and AUC scores for each galaxy subtype relative to the other three subtypes.

4.1. k-nearest Neighbor Method

The classification problem presented in this paper is relatively simple and well-suited for classical machine-learning methods that have widely available implementations: k-Nearest Neighbors, linear SVC, nonlinear SVC, etc. We use k-nearest neighbors (KNN) to establish a baseline of classification accuracy because it is the most straightforward method to use to make a classification. The classification of a source is determined by the voting results of the k neighbors who are the nearest to the input in the multidimensional parameter space.

4.1.1. KNN Performance

We use the KNN implementation in scikit-learn v0.21.2 (Pedregosa et al. 2011). The number of neighbors for voting, k, is a free hyperparameter. In Figure 4, we plot the overall accuracy as a function of k for the validation sample to determine the optimal k value, which we find to be $k = 54$. Thereafter, we trained KNN classifiers using $k = 54$. In panel (a) of Figure 5, we plot the classification accuracy for the four subtypes and the overall accuracy as a function of training sample size. The overall accuracy climbs from 0.5 to about 0.65 when the training sample size reaches 3000 and plateaus after that. The corresponding subtype accuracies are 93.4%, 59.6%, 76.0%, and 51.6% for SFGs, composites, AGNs, and LINERs. The ROC curves and AUC score for each subtype of ELG using KNN is shown in panel (b). We use 6-fold cross validation to evaluate the performance. "6-fold cross validation" means we split the training sample into six equal subsamples, and each time one subsample is used as the validation sample and the other five subsamples are used as training samples. The thin lines are the ROC curves for individual validation subsamples, and the thick lines are the mean ROC curves. The shaded areas are the 1σ errors of the mean ROC curves. KNN is quite good at distinguishing SFGs, composites, AGNs, and LINERs, with AUC scores of 0.964, 0.878, 0.860, and 0.865, respectively. We note that composites, AGNs, and LINERs have similar AUC scores, but they have very different classification accuracies in Figure 5(a). This is because the final accuracies are the result of tradeoffs among the four subtypes. KNN produces high accuracies for AGNs and composites at the expense of LINER accuracy.

4.2. Support Vector Classifier

Support vector machines have been one of the best tools for regression and classification. An SVC finds the demarcation plane by maximizing the distance between the hyperplane and the nearest point. Linear SVC assumes that the demarcation hyperplane is linear while nonlinear SVC does not make such an assumption. We use the scikit-learn implementation of nonlinear SVC for supervised learning on our sample. We also tried the linear SVC method, but it shows significantly lower accuracy for all four subtypes, thus we do not present it in this paper. The lower accuracy for linear SVC could be caused by the nonlinear demarcation lines in the 2–3 parameter BPT diagrams, which also translates to our input features.

4.2.1. SVC Performance

The classification accuracy as a function of training sample size is shown in Figure 6. The accuracy curves for the 4 subtypes stabilize after the training sample size reaches 1000 galaxies (250 per subtype). The final accuracies for SFGs, composites, AGNs, and LINERs are 92.6%, 63.8%, 79.4%, and 60.8%. The accuracy curves saturate after 1000 training sources because the separation hyperplane is optimized and more sources do not change the hyperplane significantly. With a larger training sample, the accuracy hardly changes, thus the accuracy variation is small. The ROC curves and AUC score for SFGs, composites, AGNs, and LINERs using SVC is shown in panel (b) of Figure 6. The AUC scores are 0.968 (SFGs), 0.881 (composites), 0.861 (AGNs), and 0.869 (LINERs). Despite similar AUC scores, SVC is much better than KNN for classification accuracy.

4.3. Random Forest

Another popular method for classification is decision trees. A decision tree classifies an object according to a series of criteria. However, a single tree usually introduces a cut in parameter space that is not ideal. The criteria from a single decision tree may not be ideal, so using many decision trees and letting them vote for the classification result produces much better outcomes than a single decision tree. One popular ensemble method is RF, which creates trees each with a random subset of input features and a random sample of data with replacement. We use the scikit-learn RF implementation with n_estimators = 1000, oob_score = True, and n_jobs = −1.

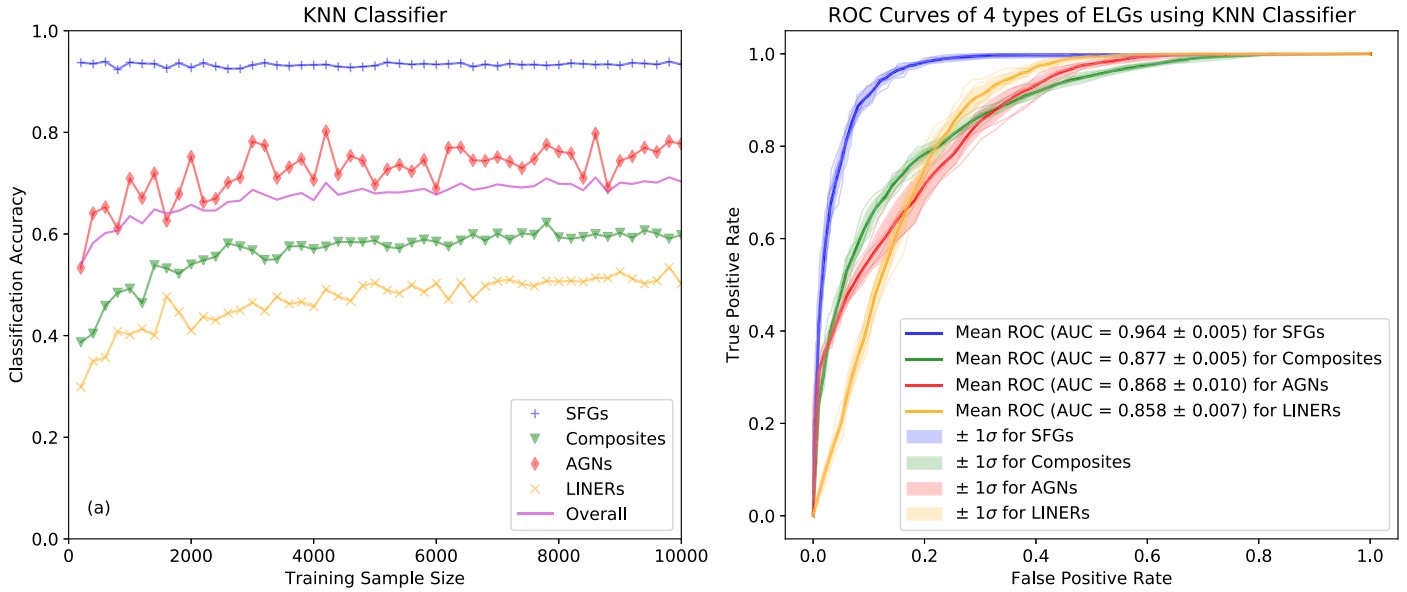

Figure 5. Panel (a): the KNN classification accuracy as a function of training sample size for the four subtypes of emission line galaxies. Blue crosses, green triangles, red diamonds, and orange x's denote SFGs, composites, AGNs, and LINERs. The magenta line shows the average classification accuracy of the four subtypes. Panel (b): ROC curves and AUC scores for each subtype of ELGs using the *k*-nearest neighbors method.

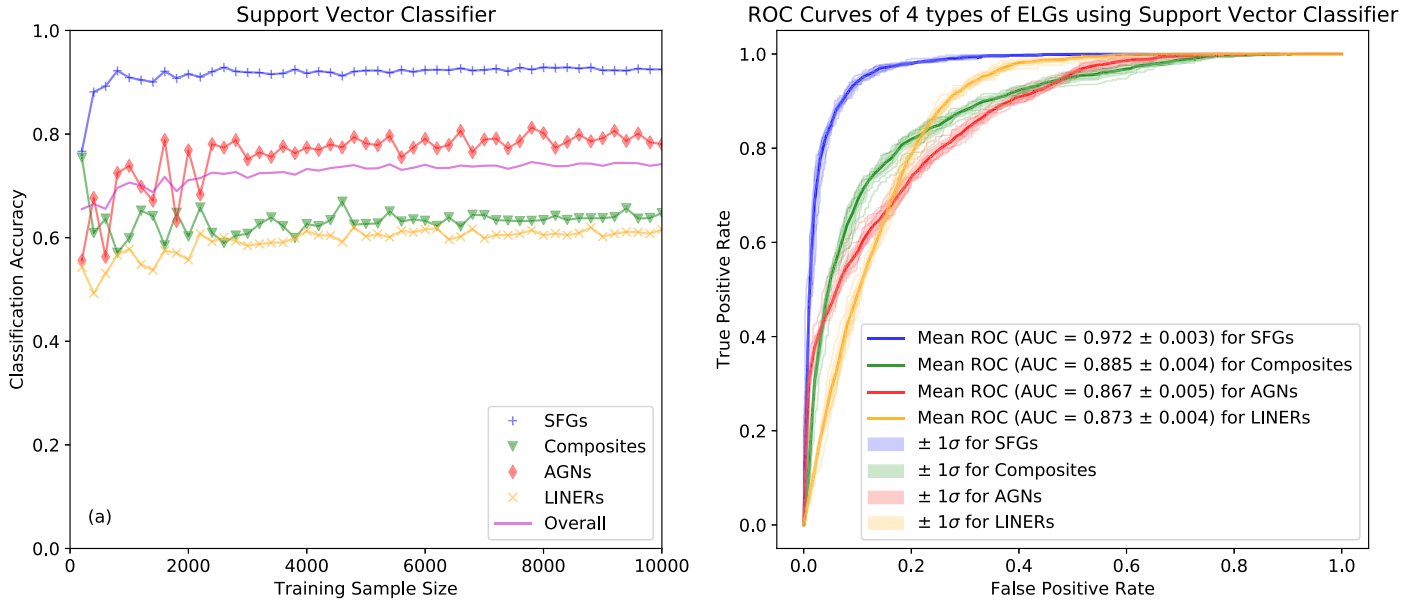

Figure 6. Panel (a): the SVC classification accuracy as a function of training sample size for the four subtypes of emission line galaxies. The legends are the same as those of Figure 5. The accuracy curves for the four subtypes are stable after the training sample size reaches 1000. Panel (b): ROC curves and AUC score for each type of ELG using SVC.

4.3.1. RF Performance

The classification accuracy as a function of training sample size is shown in Figure 7. With a very small sample size of about only 100 sources, the RF classifier gives a good overall accuracy of ∼0.65. The accuracy keeps climbing with increasing sample size and stabilizes at 0.75 at 10,000 training sources. The final accuracy for SFGs, composites, AGNs, and LINERs are 93.4%, 69.4%, 71.8%, and 65.7%, respectively. The AUC scores for SFGs, composites, AGNs, and LINERs are 0.985, 0.966, 0.876, and 0.897, respectively—a big improvement over the KNN and SVC methods.

4.4. Importance of Individual Input Parameters

Feature importance measures the distinguishing power of each feature, and opens the possibility of dropping the least important features without significantly sacrificing performance. A benefit of using gradient boosted methods is that it is straightforward to calculate importance scores for each attribute after the boosted trees are constructed. For a single decision tree, the importance is calculated by multiplying the amount that each attribute split point improves the performance measure by the number of observations for which the node is responsible. The final feature importances are the average values over all decision trees within the model. Generally, the

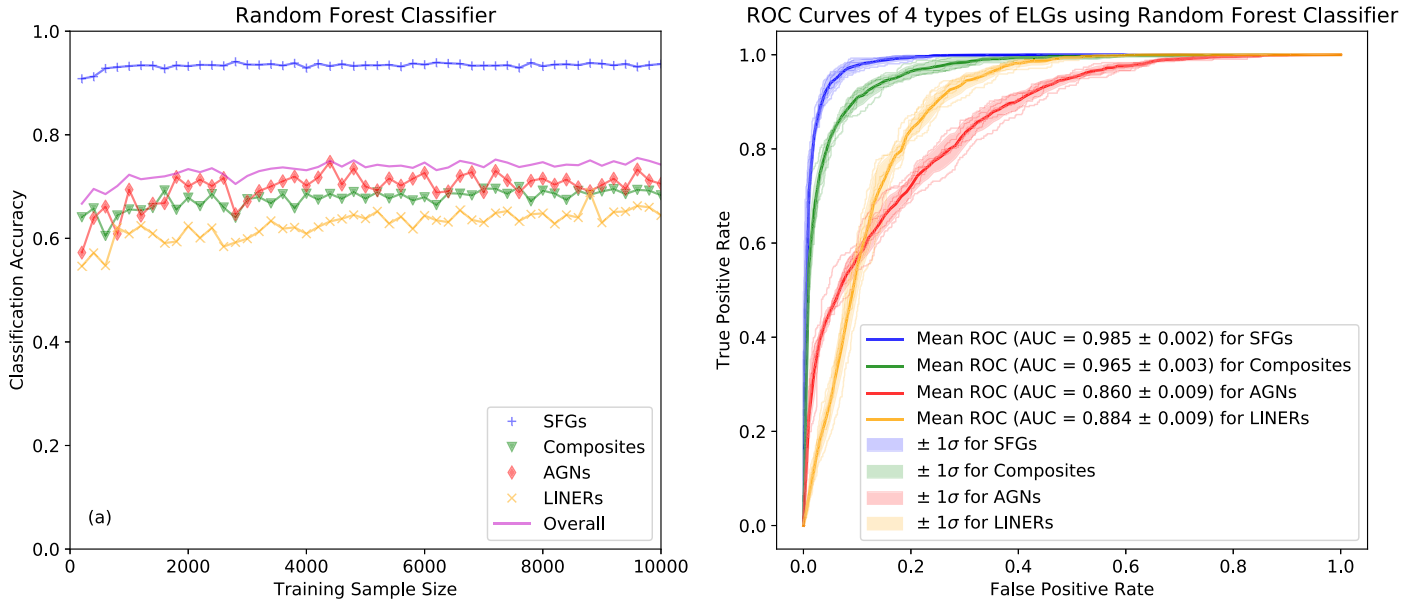

Figure 7. Panel (a): the Random Forest classifier accuracy as a function of training sample size for the four subtypes of emission line galaxies. The legends are the same as those of Figure 5. The accuracy curves for the four subtypes are stable after the training sample size reaches 10,000 galaxies. Panel (b): ROC curves and AUC scores for each subtype of ELG using the Random Forest method.

more an attribute is used to make key decisions with decision trees, the higher its relative importance.[14]

We show the importances of the eight features in Figure 8. The top three most important features are [O III]/Hβ, $\sigma_{[\mathrm{O\ III}]}$, and $g - r$. [O II]/Hβ, $u - g$, and σ_* are ranked 4–6. $r - i$ and $i - z$ are the least important with importance values about one-third of the most important feature ([O III]/Hβ). These results are impressive given that the machine-learning methods were not provided any physics principles but rather tell us what features are the most important purely from the data. Calculating feature importance can be extremely helpful for data sets with huge numbers of features whose physical meanings and connections are not well understood.

4.5. Neural Network

Finally, we apply a neural network to our classification problem. A neural network is a combination of layers of neurons, just like our brain. The parameters of a neuron (weight and bias for a linear neuron) can be adjusted through the learning process. A loss function is defined to quantify how poorly the model is at making a prediction. To improve the prediction accuracy of the model, the prediction error is back-propagated through the network, and the model is updated accordingly. The model predictions improve with additional data.

4.5.1. Neural Network Setup

Usually, the deeper the network, the better its performance. A two layer convolutional neural network is powerful enough to achieve 98% accuracy in classifying hand-written digits (0–9) from the MNIST data set. Our emission line galaxy classification problem is less complex than the MNIST task, so we use the multilayer perceptron (MLP) classifier

[14] For details on how importance in decision tree is calculated see Section 10.13.1 "Relative Importance of Predictor Variables" of the book "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," page 367.

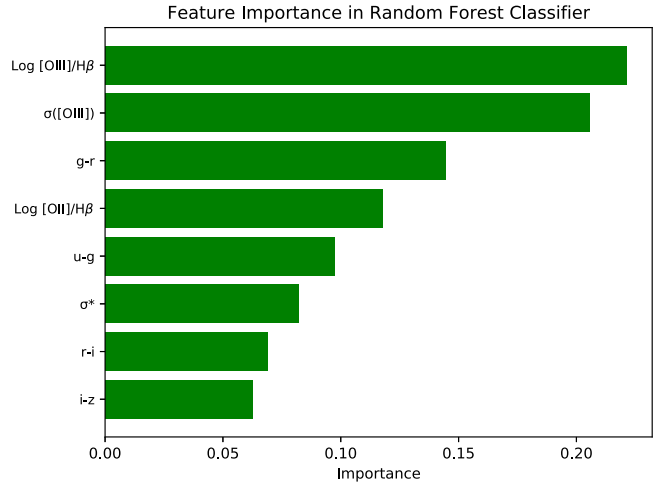

Figure 8. Importances of the eight features for the random forest classifier. The top three most important features are [O III]/Hβ, $\sigma_{[\mathrm{O\ III}]}$, and $g - r$. [O II]/Hβ, $u - g$, and σ_* are ranked 4–6. $r - i$ and $i - z$ are the least important.

implementation in the scikit-learn neural_network package with the L2 penalty parameter alpha set to 0. Other parameters are set to the default values: one hidden layer with 100 neurons and rectified linear unit function (RELU) as the activation function. The learning rate is 0.001, and maximum number of iterations is set to 200. Our model optimizes the log-loss function using the stochastic gradient descent method by setting the solver to "adam" (Kingma & Ba 2015). We set the exponential decay rate for estimates of first and second moment vector in "adam" to 0.9 and 0.999, respectively.

4.5.2. MLP Classifier Performance

The classification accuracy as a function of training sample size is shown in Figure 9. With increasing sample size, the accuracy keeps climbing and stabilizes at 0.75 at 2000 training sources. The final accuracy for SFGs, composites, AGNs, and LINERs are 93.7%, 68.8%, 76.8%, and 61.1%, respectively.

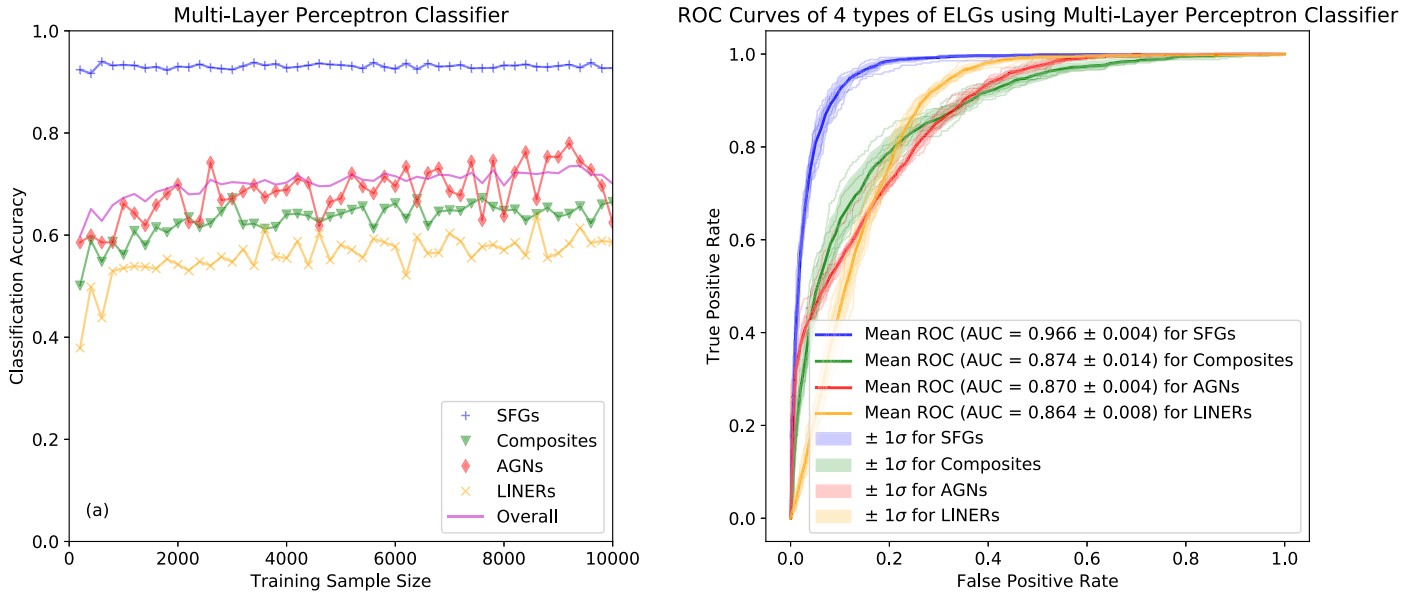

Figure 9. Panel (a): the MLP classifier classification accuracy as a function of training sample size for the four subtypes of emission line galaxies. Legends are the same as those of Figure 5. The accuracy curves for the four subtypes are stable after the training sample size reaches 2000. Panel (b): ROC curves and AUC scores for each subtype of ELGs using the multilayer perceptron classifier.

Table 1
The AUC Scores for Different Machine-learning Classifiers

Type (1)	Star-forming Galaxies (2)	Composites (3)	AGNs (4)	LINERs (5)	Average (6)
KNN	0.964 ± 0.004	0.878 ± 0.010	0.860 ± 0.007	0.865 ± 0.008	0.892
SVC	0.968 ± 0.005	0.881 ± 0.014	0.861 ± 0.009	0.869 ± 0.009	0.895
MLP	0.973 ± 0.003	0.896 ± 0.009	0.880 ± 0.004	0.877 ± 0.009	0.906
RF	0.985 ± 0.001	0.966 ± 0.004	0.876 ± 0.008	0.897 ± 0.004	0.931
RF (4 features)	0.981 ± 0.002	0.952 ± 0.003	0.870 ± 0.009	0.890 ± 0.007	0.923
RF (2 features)	0.977 ± 0.003	0.936 ± 0.010	0.865 ± 0.007	0.883 ± 0.006	0.915

Note. Columns: (1) Classifier type. (2) Star-forming galaxies AUC score. (3) Composite galaxies classification AUC score. (4) LINERs classification AUC score. (5) AGNs classification AUC score. (6) Average AUC score.

Table 2
Accuracies for Different Machine-learning Classifiers

Type (1)	Star-forming Galaxies (2)	Composites (3)	AGNs (4)	LINERs (5)	Overall (6)
KNN	0.934 ± 0.001	0.596 ± 0.006	0.760 ± 0.008	0.516 ± 0.009	0.702
SVC	0.926 ± 0.003	0.638 ± 0.008	0.794 ± 0.004	0.608 ± 0.007	0.741
MLP	0.937 ± 0.004	0.688 ± 0.017	0.768 ± 0.022	0.611 ± 0.019	0.750
RF	0.934 ± 0.001	0.694 ± 0.005	0.718 ± 0.011	0.657 ± 0.010	0.751
RF (4 features)	0.923 ± 0.002	0.637 ± 0.008	0.673 ± 0.005	0.608 ± 0.009	0.710
RF (2 features)	0.909 ± 0.002	0.594 ± 0.007	0.564 ± 0.013	0.482 ± 0.013	0.637

Note. Columns: (1) Classifier type. (2) Star-forming galaxies classification accuracy. (3) Composite galaxies classification accuracy. (4) LINERs classification accuracy. (5) AGNs classification accuracy. (6) Overall classification accuracy.

The AUC scores for SFGs, composites, AGNs, and LINERs are 0.964, 0.874, 0.867, and 0.864, respectively, which is very close to the RF classifier.

4.6. Performance Comparison

The AUC scores and accuracies for the four subtypes are given in Tables 1 and 2. The ROC curves and AUC scores for the four subtypes of ELGs for each ML method is shown in Figure 10. The average AUC score for the four subtypes are 0.892, 0.895, 0.931, and 0.892 for KNN, SVC, RF, and MLP, respectively. The rank in average accuracy is the same as the rank of AUC scores, 70.2%, 74.1%, 75.1%, and 75.0% for KNN, SVC, RF, and MLP. However, the different methods settle on different demarcation hyperplanes, resulting in different preferences. MLP has the highest accuracy for SFGs, and SVC has the highest accuracy for AGNs. RF achieves the highest accuracies for composites and LINERs. These differences reflect the different tradeoffs of the four algorithms. For robustness and performance stability, we favor the RF Classifier as the optimal method. The confusion matrix of the

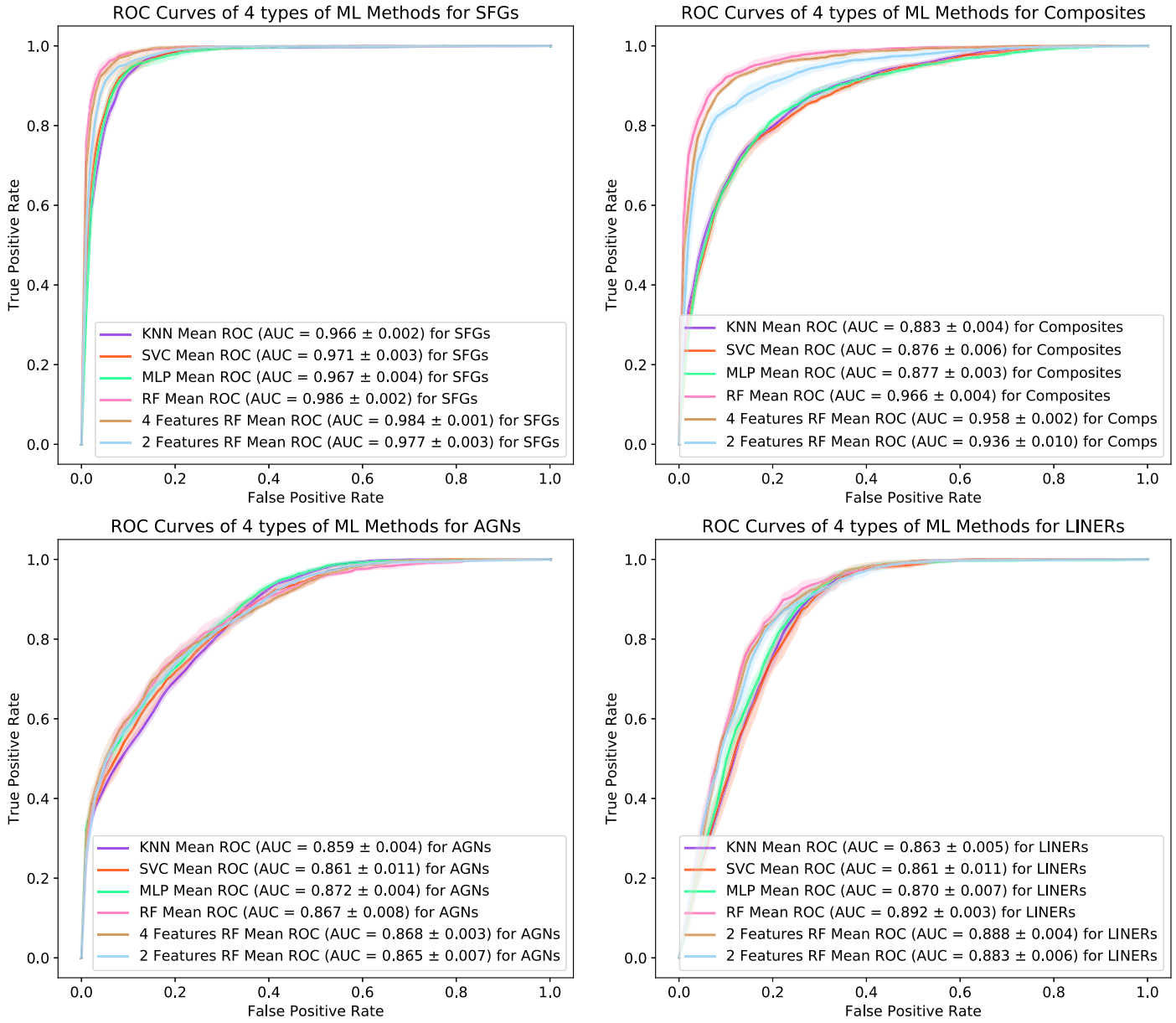

Figure 10. Comparison of ROC curves and AUC scores for each subtype of ELGs with the four ML methods. The random forest classifier achieves significantly higher AUC scores for SFGs, Composites, and LINERs than the other methods, and it performs similarly to the other methods in classifying AGNs.

RF classifier is given in Table 3. SFGs are unlikely to be confused with the other subtypes. Composites have a 23.8% probability to be confused with SFGs. AGNs have 18.8%, 8.9%, and 1.8% probabilities to be classified as LINERs, composites, and SFGs, respectively. LINERs are most likely to be misclassified as composites with 28.4% probability, and 6.7% and 0.4% probabilities to be misclassified as AGNs and SFGs, respectively.

4.7. Classification Using Only Spectroscopic Features

It is useful to construct a classifier based solely on spectroscopic features because imaging data is not always available. We reduce the feature set to [O III]/Hβ, [O II]/Hβ, σ([O III])), and stellar velocity dispersion (σ_*), and use the same training sample as in Section 4. We use an RF classifier to see how well it performs with four features compared to eight features. The accuracy curves and ROC curves are given in

Table 3
Confusion Matrix for the Random Forest Classifier

Type	Star-forming Galaxies	Composites	AGNs	LINERs
(1)	(2)	(3)	(4)	(5)
True SFGs	0.937	0.060	0.003	0.000
True Composites	0.238	0.683	0.018	0.061
True AGNs	0.018	0.089	0.705	0.188
True LINERs	0.004	0.284	0.067	0.644

Note. Confusion matrix of the random forest classifier. Each row gives the probabilities that galaxies of a given subtype of ELG are classified as each of the four subtypes. Star-forming galaxies are unlikely to confuse with the other subtypes. Composites have a 23.8% probability to be confused with star-forming galaxies. AGNs have 18.8%, 8.9%, and 1.8% probabilities to be classified as LINERs, composites, and star-forming galaxies. LINERs are most likely to be misclassified as composites with 28.4% probability, with 6.7% and 0.4% probabilities to be misclassified as AGNs and star-forming galaxies.

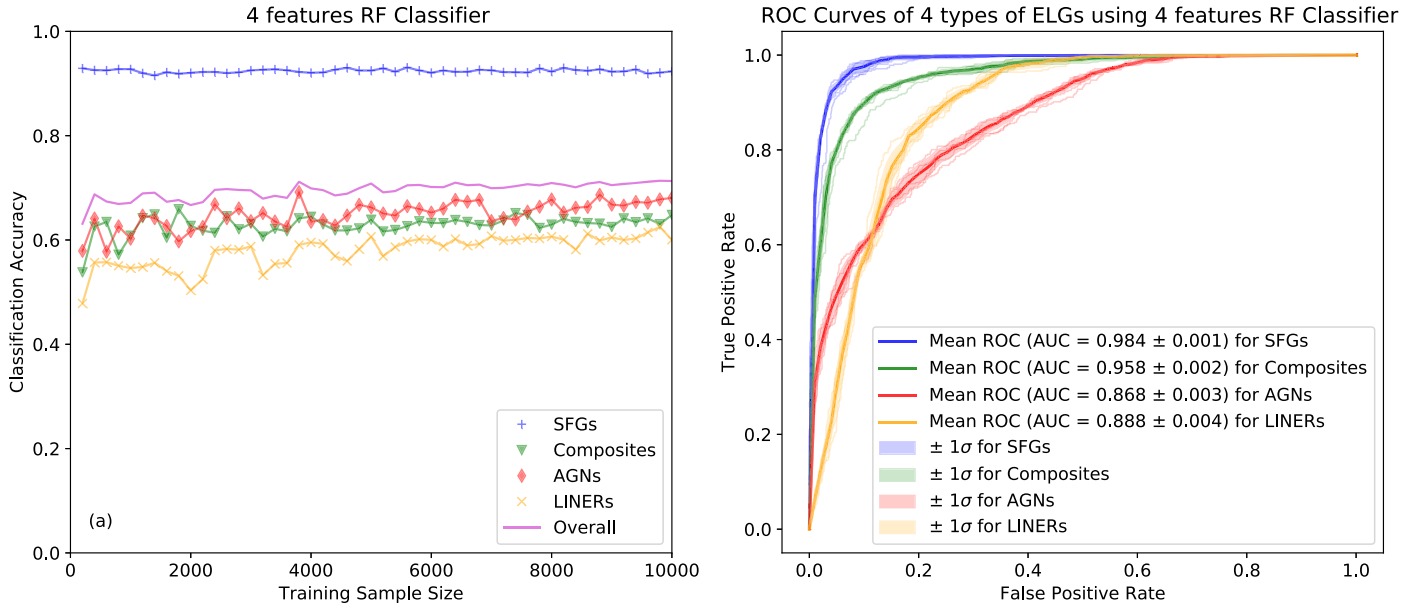

Figure 11. Panel (a): the random forest classifier accuracy as a function of training sample size for four subtypes of emission line galaxies, using only four features: [O III]/Hβ, [O II]/Hβ, σ([O III])), and stellar velocity dispersion. Legends are the same as Figure 5. The accuracy curves for four subtypes are stable after the training sample size reaches 10,000 galaxies. Panel (b): ROC curves and AUC scores for each type of ELGs using the random forest method and only four spectroscopic features.

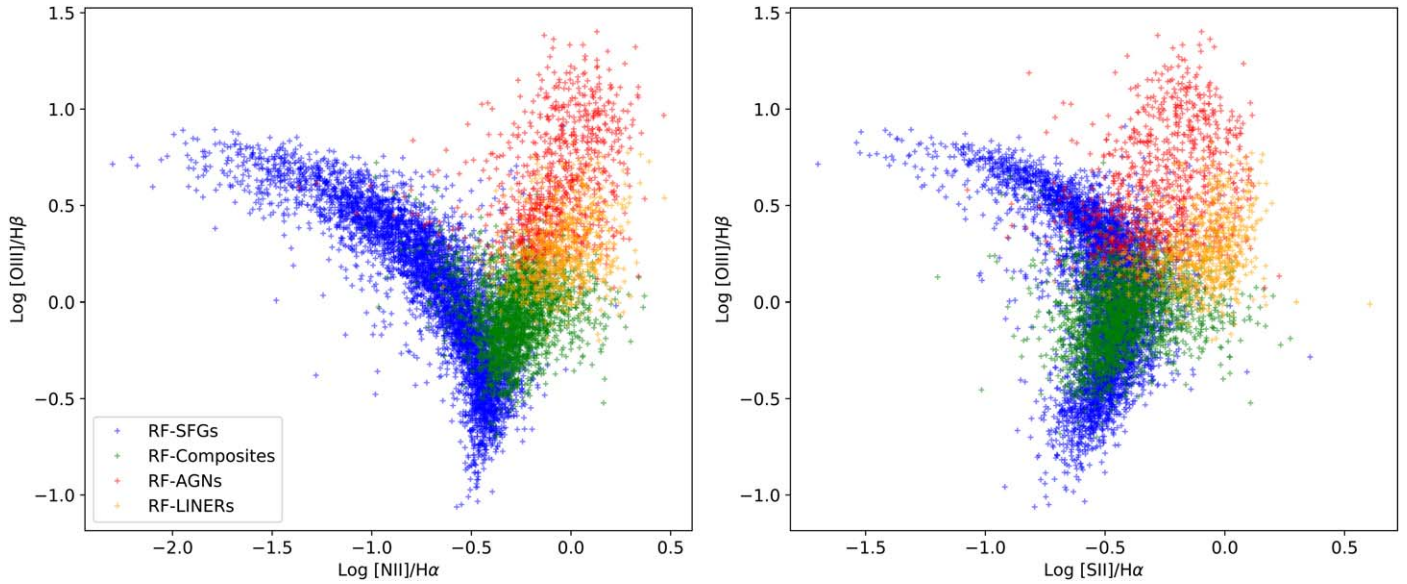

Figure 12. RF-classified $z < 0.32$ galaxies of four subtypes on the BPT diagram. The RF classifier does an excellent job of reproducing the BPT diagram classification.

Figure 11. The AUC scores drop from 0.985, 0.966, 0.876, and 0.897 to 0.981, 0.952, 0.870 and 0.890 for SFGs, composites, AGNs, and LINERs, respectively. Thus, the colors do help in classification, but reducing the feature set to only spectroscopic features does not degrade the classification performance significantly. It would be ideal to have 8 features for the classification ELGs, but using only four spectroscopic features [O III]/Hβ, [O II]/Hβ, σ([O III]) and stellar velocity dispersion (σ_*) can give a very similar result.

4.8. Machine-learning Classifications on the BPT, Kinematic–Excitation, and Mass–Excitation Diagrams

In Figure 12, we show the RF-classified $z < 0.32$ galaxies on the BPT diagram. The RF classifier reproduces the BPT diagram classification well, so Figures 1 and 12 appear to be very similar. In Figure 13, we plot the RF-classified $0.32 < z < 0.8$ galaxies on the kinematic–excitation (KEx; Zhang & Hao 2018) and mass–excitation (MEx; Juneau et al. 2011) diagrams. The stellar mass is drawn from the SDSS Galaxy Properties from the Wisconsin Group value-added catalog[15] (Chen et al. 2012). We chose the stellar mass derived

[15] https://www.sdss.org/dr12/spectro/galaxy_wisconsin/

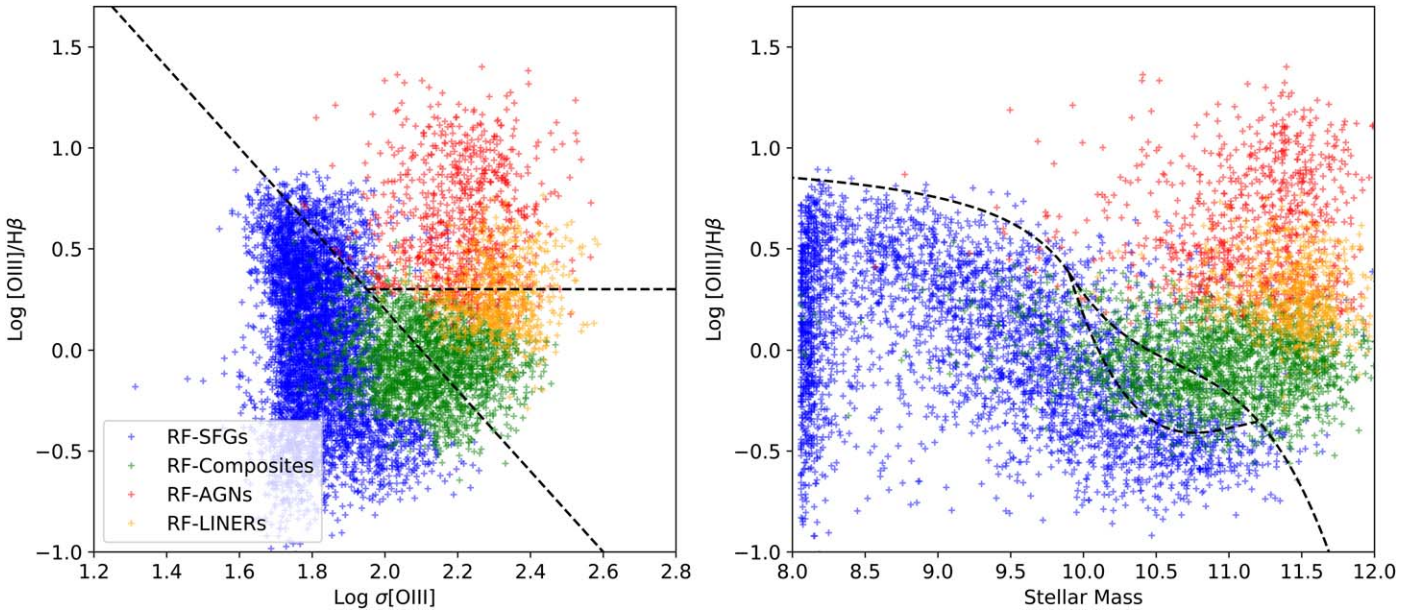

Figure 13. RF-classified $0.32 < z < 0.8$ galaxies on the kinematic–excitation (KEx; left panel; Zhang & Hao 2018) and mass–excitation (MEx; right panel; Juneau et al. 2011) diagrams. The RF classification results are consistent with the demarcation lines proposed in those two works.

using the Maraston & Strömbäck (2011) templates. The RF classification results are quite consistent with the KEx and MEx demarcation lines from Zhang & Hao (2018) and Juneau et al. (2011), respectively. In terms of accuracy, the KEx diagram gives classification accuracies of 89%, 75.6%, and 81% for SFGs, composites, and AGNs, respectively. The classification accuracies using the MEx diagram are 94.4%, 47.8%, and 54% for SFGs, composites, and AGNs, respectively. By comparison, the accuracies of the RF classification for SFGs are 93.4%, 69.4%, and 71.8%. The KEx diagram achieves very good accuracy for composites and AGNs by sacrificing the accuracy of SFGs. To make an apples-to-apples comparison, we derive the ROC curves and AUC scores using the RF classifier but restrict it to only use the same two features as the KEx diagram: [O III]/Hβ and σ([O III]). The AUC scores drop from 0.985, 0.966, 0.876, and 0.897 to 0.977, 0.936, 0.865, and 0.883 for SFGs, composites, AGNs, and LINERs, respectively. The accuracies of the 2 feature RF classifier (see Table 2) are significantly lower than that of the 8 feature RF. Our results indicate that the RF classifier using 8 features gives as consistent of a classification as the BPT diagram, and it outperforms the KEx diagram and MEx diagram.

5. Applying the RF Classifier to $0.32 < z < 0.8$ Emission Line Galaxies

We apply the RF classifier trained in Section 4.3 to 49,272 $0.32 < z < 0.8$ emission line galaxies in Section 2.3. We use the kcorrect package to k-correct the photometry to $z = 0.1$ u, g, r, i, and z magnitudes. [O III]/Hβ, [O II]/Hβ, $\sigma_{[\mathrm{O\ III}]}$, and σ_* are measured using eBOSS spectra. The RF classifies 23,919 galaxies as star-forming, 13,536 as composites, 9448 as AGNs, and 2369 as LINERs. The ideal method to test our classifications would be to observe the galaxies with near-IR spectra to cover the rest-frame optical wavelength range so that a BPT classification is possible because this is the only way to accurately classify the sample into the four subtypes. This will be possible when spectra from surveys like MOSFIRE Deep Evolution Field Survey (Kriek et al. 2015; Sanders et al. 2016) become publicly available. In lieu of large numbers of near-IR spectra, we compare the stacked spectra of the four subtypes using the stacking code developed in Comparat et al. (2016).[16] The stacked spectrum of each of the four subtypes should be significantly different from each other, and they should be generally consistent with their $z < 0.32$ counterparts.

Figure 14 shows the rest-frame 3400–5050 Å stacked spectra of RF-classified $0.32 < z < 0.8$ SFGs, composites, AGNs, and LINERs and their $z < 0.32$ counterparts classified using the BPT diagrams. SFGs show prominent absorption features and the least steep continuum. AGNs show a steeper continuum than SFGs and prominent emission lines. Composites have features in between those of SFGs and AGNs, as expected. LINERs show the steepest spectrum and also significant emission lines. The spectral shape of RF-classified sources are highly consistent with BPT-classified low redshift ELGs of the same subtype. The [O III]/Hβ ratios are consistent with expectations, too. This strongly suggests that the RF classifier is correctly classifying the four subtypes of ELGs.

Despite these consistencies, there are noticeable differences between the RF-classified and BPT-classified samples. The equivalent width of RF-classified composites and LINERs are significantly higher than their $z < 0.32$ BPT-classified counterparts. There are at least two reasons for this difference. First, the sample selection criteria require that the signal-to-noise ratios of Hβ and [O III] are greater than 3, effectively selecting stronger emission line galaxies at intermediate redshift than at low redshift. Second, the intermediate redshift composite and LINER groups are contaminated by AGNs and SFGs, whose much stronger emission lines will bias stacked spectra. This contamination does not dominate the spectra because the [O III]/Hβ ratios are consistent with the low redshift values. Thus, we conclude that the selection effect is the main reason for the difference in the equivalent widths of the low and intermediate redshift composites and LINERs.

[16] https://github.com/JohanComparat/pySU/blob/master/galaxy/python/SpectraStackingEBOSS.py

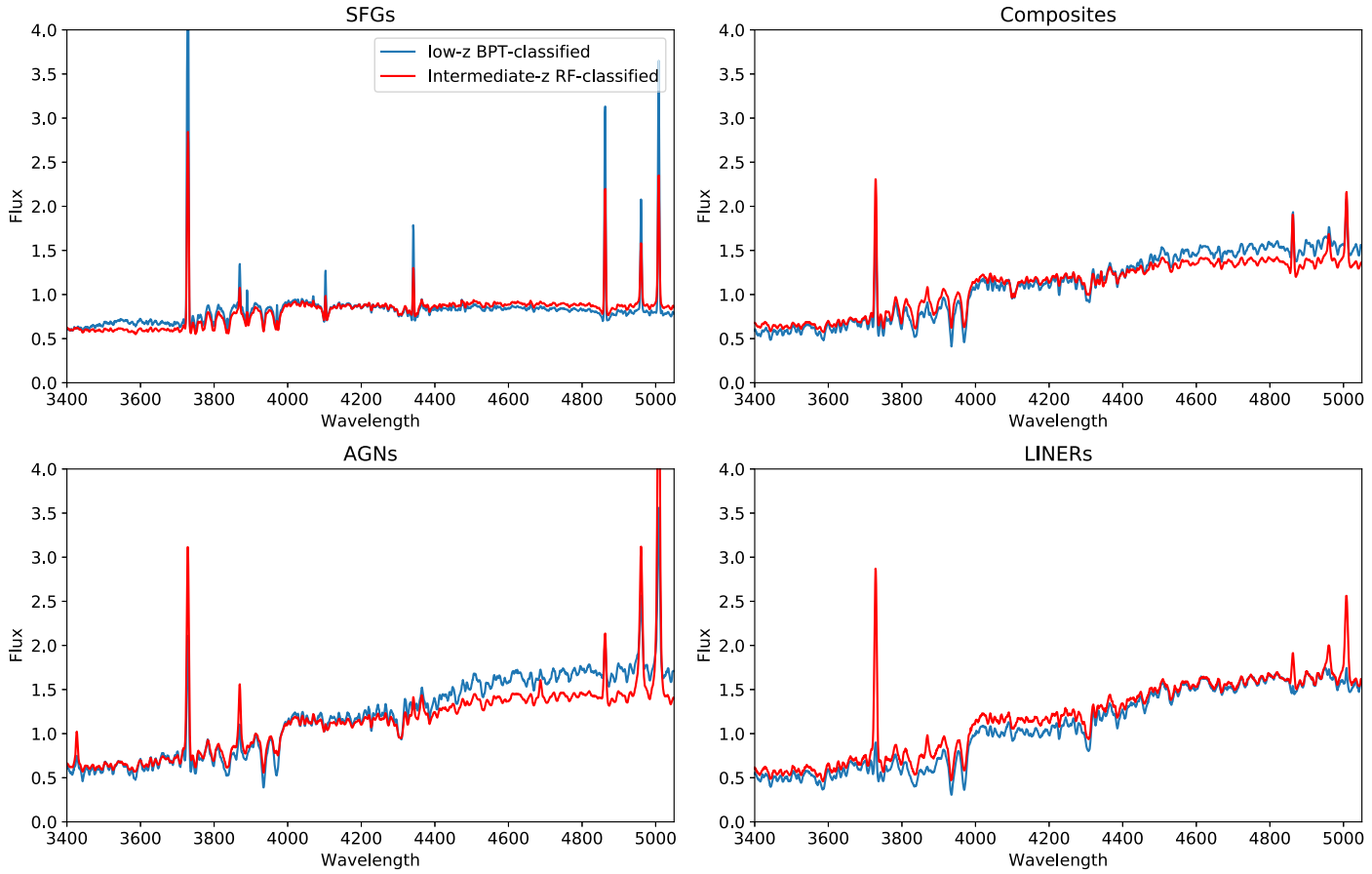

Figure 14. Comparison of stacked spectra of $0.32 < z < 0.8$ SFGs, composites, AGNs, and LINERs classified using the Random Forest classifier described in Section 4.3 shown in red and the BPT-classified $z < 0.32$ SFGs, composites, AGNs, and LINERs shown in blue.

6. Conclusions

In this paper, we consider the classification of intermediate redshift emission line galaxies using supervised machine-learning classification algorithms. We use measurements available for optical spectra of galaxies at $z < 0.8$: [O III]/Hβ, [O II]/Hβ, [O III] line width (σ([O III])), stellar velocity dispersion (σ_*), $u - g$, $g - r$, $r - i$, and $i - z$ color as input. A $z < 0.3$ emission line galaxy sample classified and labeled using standard optical diagnostic diagrams is selected as a training sample. We use k-nearest neighbors (KNN), SVC, RF, and a MLP-NN to train models that predict which class a galaxy belongs to given a set of input. ROC curve and area under curve (AUC) scores are used to quantify the distinguishing power of the different classifiers. RF has the best AUC score for classifications of all four subtypes, while the relative ranking of the other three algorithms in both AUC scores and accuracies is MLP > SVC > KNN. The RF classification accuracies are 93.4%, 69.4%, 71.8%, and 65.7% for SFGs, composites, AGNs, and LINERs, respectively. The three most important features are [O III]/Hβ, $\sigma_{[\mathrm{O\ III}]}$, and $g - r$. Reducing the input to the four spectroscopic features results in slightly degraded accuracies of 92.3%, 63.7%, 67.3%, and 60.8%. The stacked spectra of the four subtypes classified using the RF model are consistent with the stacked spectra of low redshift BPT-classified ELGs. The machine-learning classification tool will play an important role in emission line galaxy physics in upcoming large sky surveys like DESI, PFS, and 4MOST.

K.Z. thanks Xiaosheng Huang for a helpful discussion on machine-learning techniques. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS- IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS website is www.sdss.org.

SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU)/University of Tokyo, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatory of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional/MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of

Washington, University of Wisconsin, Vanderbilt University, and Yale University.

ORCID iDs

Kai Zhang https://orcid.org/0000-0002-9808-3646
David J. Schlegel https://orcid.org/0000-0002-5042-5088
Brett H. Andrews https://orcid.org/0000-0001-8085-5890
Renbin Yan https://orcid.org/0000-0003-1025-1711

References

Acquaviva, V. 2019, arXiv:1901.05978
Aguado, D. S., Ahumada, R., Almeida, A., et al. 2019, ApJS, 240, 23
Azadi, M., Coil, A. L., Aird, J., et al. 2017, ApJ, 835, 27
Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, PASP, 93, 5
Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28
Blanton, M. R., & Roweis, S. 2007, AJ, 133, 734
Boyle, B. J., Shanks, T., Croom, S. M., et al. 2000, MNRAS, 317, 1014
Bradley, A P 1997, Pattern Recognition, 30, 1145
Chen, Y.-M., Kauffmann, G., Tremonti, C. A., et al. 2012, MNRAS, 421, 314
Comparat, J., Zhu, G., Gonzalez-Perez, V., et al. 2016, MNRAS, 461, 1076
Dawson, K. S., Kneib, J.-P., Percival, W. J., et al. 2016, AJ, 151, 44
de Jong, R. S., Bellido-Tirado, O., Chiappini, C., et al. 2012, Proc. SPIE, 8446, 84460T
de la Calleja, J., & Fuentes, O. 2004, MNRAS, 349, 87
Dey, A., Schlegel, D. J., Lang, D., et al. 2019, AJ, 157, 168
Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441
Fawcett, T 2006, PaReL, 27, 861
Ferri, C., Hernández-Orallo, J., & Modroiu, R. 2009, PaReL, 30, 27
Hand, D. J., & Till, R. J. 2001, Machine Learning, 45, 171
Hocking, A., Geach, J. E., Sun, Y., et al. 2018, MNRAS, 473, 1108
Huang, X., Domingo, M., Pilon, A., et al. 2019, arXiv:1906.00970
Jacobs, C., Collett, T., Glazebrook, K., et al. 2019a, MNRAS, 484, 5330
Jacobs, C., Collett, T., Glazebrook, K., et al. 2019b, arXiv:1905.10522
Jacobs, C., Glazebrook, K., Collett, T., et al. 2017, MNRAS, 471, 167
Juneau, S., Dickinson, M., Alexander, D. M., & Salim, S. 2011, ApJ, 736, 104
Juneau, S, Dickinson, M, Bournaud, F, et al. 2013, ApJ, 764, 176
Kauffmann, G., Heckman, T. M., Tremonti, C., et al. 2003, MNRAS, 346, 1055
Kewley, L. J., Dopita, M. A., Leitherer, C., et al. 2013a, ApJ, 774, 100
Kewley, L. J., Groves, B., Kauffmann, G., & Heckman, T. 2006, MNRAS, 372, 961
Kewley, L. J., Maier, C., Yabe, K., et al. 2013b, ApJL, 774, L10
Kingma, D. P., & Ba, J. 2015, 3rd Int. Conf. for Learning Representations arXiv:1412.6980
Kriek, M., Shapley, A. E., Reddy, N. A., et al. 2015, ApJS, 218, 15
Lamareille, F. 2010, A&A, 509, A53
Levi, M., Bebek, C., Beers, T., et al. 2013, arXiv:1308.0847
Lintott, C., Schawinski, K., Bamford, S., et al. 2011, MNRAS, 410, 166
Maraston, C., & Strömbäck, G. 2011, MNRAS, 418, 2785
Marocco, J., Hache, E., & Lamareille, F. 2011, A&A, 531, A71
Metcalf, R. B., Croft, R. A. C., & Romeo, A. 2018, MNRAS, 477, 2841
Metz, C E 1978, Seminars in Nuclear Medicine, 8, 283
Mossman, D 1999, Medical Decision Making, 19, 78
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Petrillo, C. E., Tortora, C., Chatterjee, S., et al. 2017, MNRAS, 472, 1129
Pourrahmani, M., Nayyeri, H., & Cooray, A. 2018, ApJ, 856, 68
Rola, C. S., Terlevich, E., & Terlevich, R. J. 1997, MNRAS, 289, 419
Sanders, R. L., Shapley, A. E., Kriek, M., et al. 2016, ApJ, 816, 23
Srinivasan, A. 1999, Note on the Locations of Optimal Classifiers in N-Dimensional ROC Space Technical Report, PRG-TR-2-99, Oxford University Computing Laboratory
Stasińska, G., Cid Fernandes, R., Mateus, A., Sodré, L., & Asari, N. V. 2006, MNRAS, 371, 972
Takada, M., Ellis, R. S., Chiba, M., et al. 2014, PASJ, 66, R1
Tamura, N., Takato, N., Shimono, A., et al. 2016, Proc. SPIE, 9908, 99081M
Tresse, L., Rola, C., Hammer, F., et al. 1996, MNRAS, 281, 847
Trouille, L., Barger, A. J., & Tremonti, C. 2011, ApJ, 742, 46
Trump, J. R., Konidaris, N. P., Barro, G., et al. 2013, ApJL, 763, L6
Veilleux, S., & Osterbrock, D. E. 1987, ApJS, 63, 295
Weiner, B. J., Willmer, C. N. A., Faber, S. M., et al. 2006, ApJ, 653, 1027
Yan, R., Ho, L. C., Newman, J. A., et al. 2011, ApJ, 728, 38
York, D. G., Adelman, J., Anderson, J. E., Jr., et al. 2000, AJ, 120, 1579
Zhang, K., & Hao, L. 2018, ApJ, 856, 171