Galactic activity diagnostics based on WISE photometry and machine learning methods

school of sciences and engineering

Department of Physics

University of Crete

Undergraduate degree thesis

by

Charalampos Daoutis

Supervisor

Prof. Andreas Zezas

2021

Abstract

One of the most important and difficult problems in modern Astrophysics is the classification of galaxies based on their activity. A lot of progress has been done over the years with numerous diagnostics to have been developed; optical and infrared methods being the most successful and popular among them. In the recent years with the advent of the all-sky surveys from space telescopes, infrared diagnostics for AGN selection methods have become popular. Unfortunately, we find that some of them are neither complete or reliable in galaxies located in the local Universe. In addition, the class of passive galaxies is absent from these diagnostics. For these reasons, we embarked in the development of a new three-dimensional activity diagnostic based on machine-learning methods and WISE infrared photometry. In this project, we consider the classes of star-forming, AGN, composite and passive galaxies. We find that a diagnostic based on three features derived from the three WISE bands (1, 2 and 3): absolute magnitude on the band 2, band 1 – band 2 color and band 2 – band 3 color, offers precision above 90% for star-forming and passive galaxies. In addition, using the new diagnostic, we classify 68.7% of the galaxies found in the HECATE catalog. Finally, we estimate the activity demographics in the local Universe based on the results from the classification of the full HECATE catalog.

Contents

1 Introduction

Galaxies can be classified into different categories based on their activity. Some form new stars (star-forming galaxies, also referred as HII due to their HII-region like spectra), other ones can present intense nuclear activity fueled by their supermassive nuclear black hole (Active Galactic Nuclei or AGN). In addition, galaxies can present, simultaneously, both of these behaviors. These are known as composite galaxies or Transition Objects (e.g. Ho et al. 1997). In the fourth galactic category we find galaxies that host old stellar populations and contain low amounts of gas or dust. These are the passive galaxies. Finally, there is also the LINER (Low Ionization Nuclear Emission-line region) galaxies (Heckman 1980). These galaxies can be separated into two distinct categories: LINERs which are power by a supermassive black hole and present broad emission-lines (LINER type 1, Ho et al. 1997) and LINERs type 2 which are believed to be powered by UV emission from post-AGB stars.

Until now, the best way to discriminate between the three classes of galaxies mentioned above (star-forming, AGN and composite) is through the BPT diagrams (Baldwin, Phillips & Terlevich -1981). These are 2-Dimensional diagrams that separate galaxies into HII regions (star-forming), AGN(Seyfert) LINERs and composites using the characteristic emission-line ratio fluxes of: [OIII](λ5007)/Hβ, [NII](λ6583)/Hα and [SII](λ6716, λ6731)/Hα. The diagram is a plot of [OIII](λ5007)/Hβ against either [NII](λ6583)/Hα or one of the [SII](λ6716, λ6731)/Hα and [OI](λ 6300)/Hα. The classification of the galaxy depends on the location of the galaxy in those diagrams. Although this has been a highly accurate and reliable method for galaxy activity classification purposes for many years, it presents some disadvantages. One of them is that in order to classify a galaxy one needs to obtain optical spectrum. Measurements in the visible (optical) spectrum though, are difficult to be acquired. The main problem is that obtaining spectra requires long exposures and cannot be done for the entire galactic population. A second reason is due to absorption by the interstellar medium (ISM) of our galaxy. An additional reason is that some of the emission lines, are weak emissions, making the process even more difficult. In order to overcome these difficulties, new methods for classifying galaxies have emerged using infrared photometry and more specifically in the range of mid-infrared 3-24 μm.

Some of the first infrared diagnostics were defined by Stern et al. 2005 and Donley et al. 2012 with observations based on the Spitzer Space Telescope (Werner et al. 2004). Subsequently, the launch of the WISE satellite (Wide-field Infrared Survey Explorer, Wright et al. 2010) enabled systematic studies of galaxies by providing sensitive all-sky photometry in the 3-24 μm. This led to the development of a new family of diagnostic diagrams. One widely used diagnostic based on WISE infrared photometry for AGN identification, is the simple criterion of W1-W2 ≥ 0.8 (Assef et al. 2013), where the W1 and W2 are the two WISE bands 3.4 and 4.6 μm respectively (Figure 1).

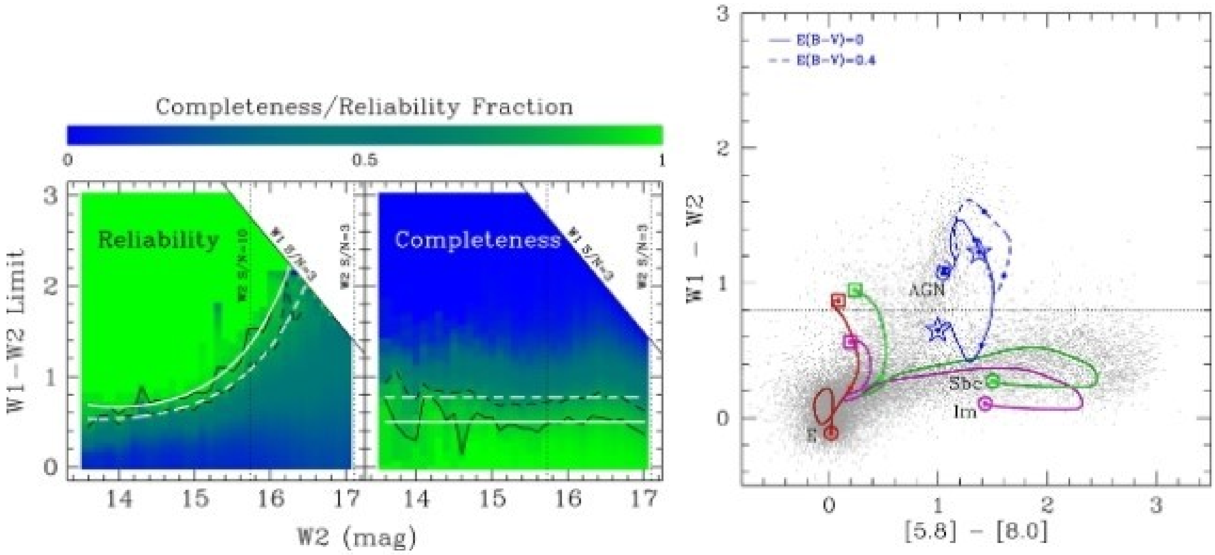


Figure 1: On the left, the reliability and completeness as function of W2 magnitude for the AGN selection criterion of W1-W2 ≥ 0.8 of Assef et al. 2013. On the right, templets for different galaxy types in the Boötes field. The templets of each galaxy type on a W1-W2 color against the [5.8] - [8.0] color diagram for different redshift values from 0 (open star) - 2 (open square) for the E (red line), Sbc (green line), and Im (magenta line). The AGN templates corresponds to solid blue line (no reddening) and dashed blue line (reddening E(B-V) = 0.4) for the differed redshifts values from 0 (open star) - 6 (open circle).

Another widely used diagnostic based on infrared colors and in the W1, W2 and W3 WISE bands in particular, is a plot W1-W2 color against the W2-W3 color. A wedge in the top right area defines the locus of AGN galaxies (Mateos et al. 2012) as seen in Figure 2.

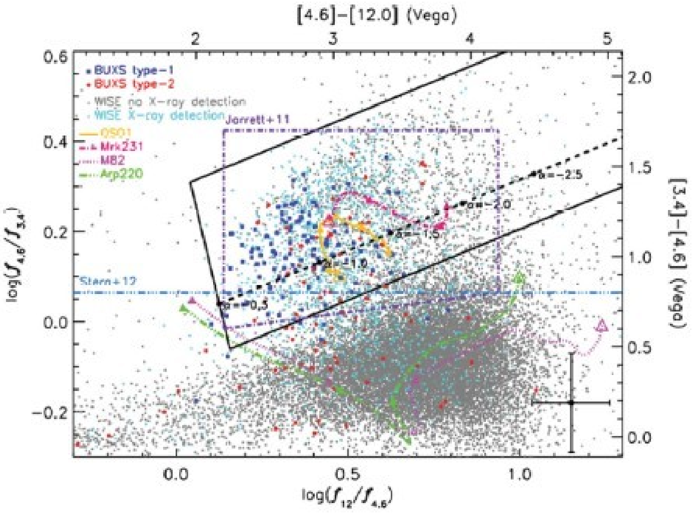


Figure 2: The AGN selection wedge (solid black lines) on W1-W2 color against W2-W3 color plot. Figure from Mateos et al. 2012.

Unfortunately, even though, these two infrared AGN selection methods have had great success in the high redshift galaxies in surveys like CANDELS (Koekemoer et al. 2011), we find that, in the local Universe, these two mid-IR classification methods are not sufficient. In a sample of galaxies taken from the Sloan Digital Sky Survey, most of the AGN galaxies are located below the W1-W2 = 0.8 line (Assef et al. 2013) or are located outside the AGN wedge (Mateos et al. 2012) as can been seen in Figures 3 and 4.

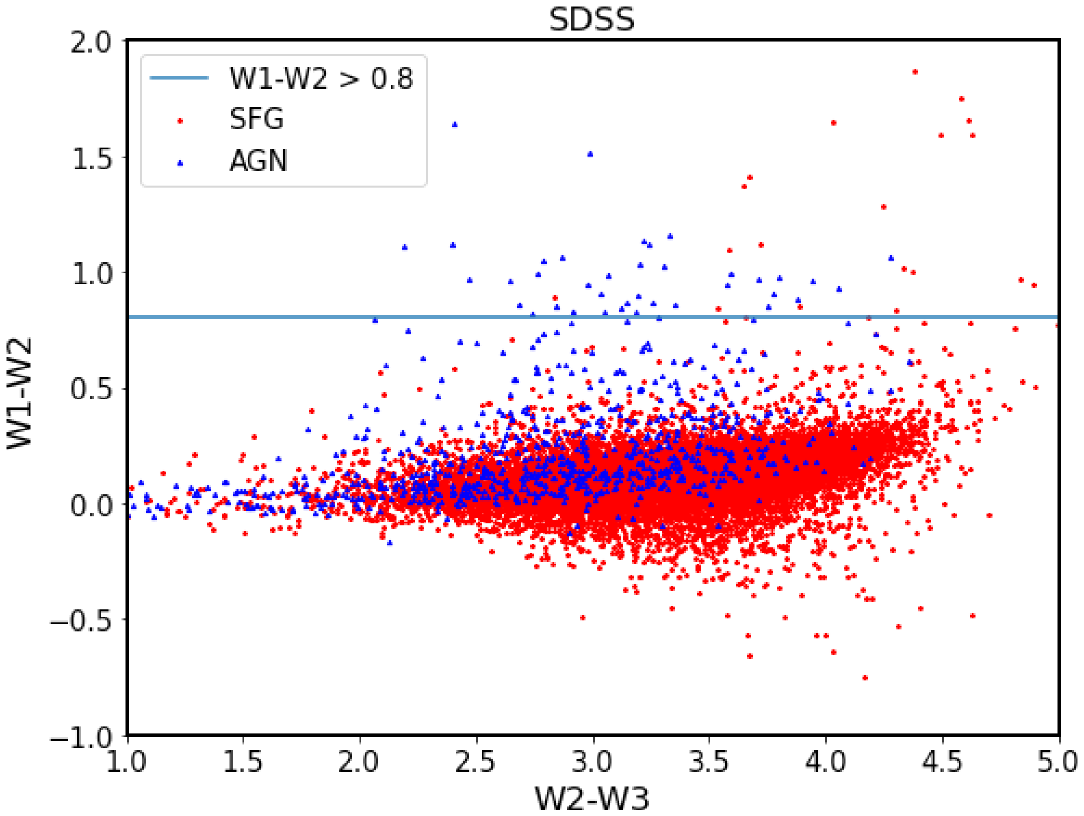


Figure 3: W1-W2 against W2-W3 plot for an SDSS subsample of galaxies within 200Mpc. Red and blue represent the two classes of star-forming and AGN respectively. These labels were defined based on the multidimensional optical-line ratio classification scheme of Stampoulis et al. 2019. The blue line represents the W1-W2 ≥ 0.8 AGN selection criterion of Assef et al. 2013. From the above plot we conclude that a significant fraction of AGN is below the selection boundary and therefore there not selected as AGN. Also, there is a lot of blending with the star-forming galaxies.

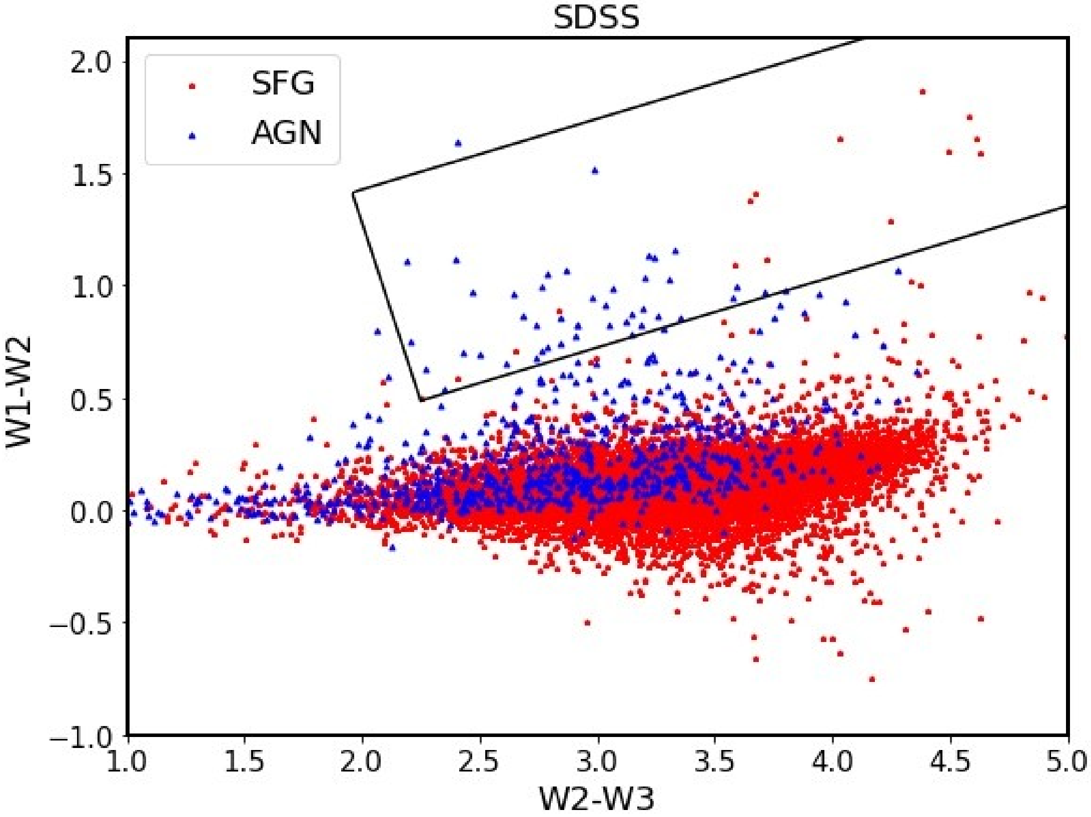


Figure 4: W1-W2 color against W2-W3 color plot for an SDSS subsample of the galaxies within 200Mpc. Red and blue represent the two classes of star-forming and AGN respectively based on the multidimensional optical-line ratio classification scheme of Stampoulis et al. 2019. The AGN wedge (black solid line) on the top right corner represents the AGN selection criterion of Mateos et al. 2012. It is obvious that a significant fraction of the AGN lie outside of the AGN selection locus.

For these reasons, in order to overcome these limitations, we embarked in the development of a new activity diagnostic based on infrared WISE photometry and machine-learning methods. We also consider in this diagnostic the class of passive galaxies. The methods used for this purpose, was Random Forest and Support Vector Machine or SVM (Cortes & Vapnik 1995) on the multi-band photometry from the WISE All-sky survey. This new diagnostic is a three-dimensional based on WISE infrared photometry. The three features used for its definition are the absolute magnitude on the band 2, the band 1 – band 2 (color) and the band 2 –band 3 (color).

2 Data Sample

2.1 The Wide-field Infrared Survey Explorer (WISE) photometry

The Wide-field Infrared Survey Explorer (WISE), is a satellite that mapped almost the entire sky. The WISE All-Sky Release Source Catalog has covered 42,195 deg^2, or 99.86% of the entire sky in four broad bands in the ~3 – 25 μm range. The bands W1, W2, W3 and W4 have effective wavelengths at 3.4, 4.6, 12, and 22 μm respectively. The angular resolution of the four bands W1, W2, W3 and W4 was 6.1, 6.4, 6.5 and 12.0 arcseconds respectively (Wright et al. 2010). The relative response in every WISE band is shown in Figure 5. WISE provides several advantages for the classification of large populations of galaxies as it is more sensitive than previous broad-band IR surveys, it covers the 3-25 μm range which includes several important diagnostic features, e.g. PAH (Polycyclic Aromatic Hydrocarbons) emission and the transition from the stellar continuum to dust emission).

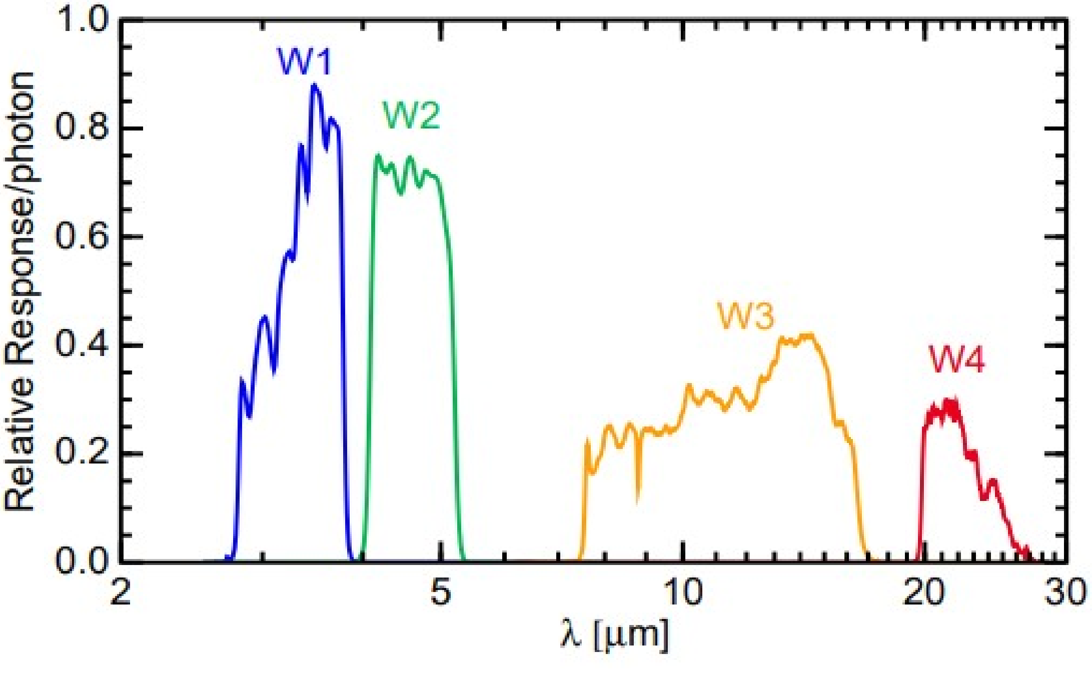


Figure 5: The relative response of the WISE bands in electrons per photon. Figure of Wright et al. 2010.

This survey offers different photometry profiles and apertures. In this project, we will use the w?gmag_2 and w?gmag (the ? corresponds to different band number 1,2,3 and 4 for W1, W2, W3 and W4 respectively). The w?mag_2 photometry is the calibrated source brightness measured within a circular aperture of 8.25 arcseconds radius centered on the source position for every WISE band and no aperture-correction has been applied. The background sky was measured with inner and outer radius of 50 and 70 arcseconds respectively. No curve of growth correction was applied. The w?gmag photometry is the magnitude of source measured in the elliptical aperture for every WISE band (the ? corresponds to 1,2,3 and 4 for W1, W2, W3 and W4 respectively). The features of the elliptical aperture (semi major axis and position angle) for the w?gmag photometry are calculated based on the 2MASS survey (Skrutskie et al. 2006). This ensures that the full extent of the galaxy is accounted for. The WISE analysis also provides extended source photometry, which however, is subject to significant photometric uncertainties due to the low signal-to-noise ratio in the lower-surface brightness regions of the galaxies.

2.2 The sample of galaxies

The sample of galaxies used in this study, originates from the HECATE catalog of galaxies (Kovlakas et al. 2021). This catalog contains galaxies located between 0 to 200 Mpc. The wealth of information and data in this catalog, makes it suitable for multi-wavelength and multi-messenger studies in the local Universe. The distribution of the galaxies in the sky map and the Venn diagram for this catalog are shown in Figures 6 and 7 respectively.

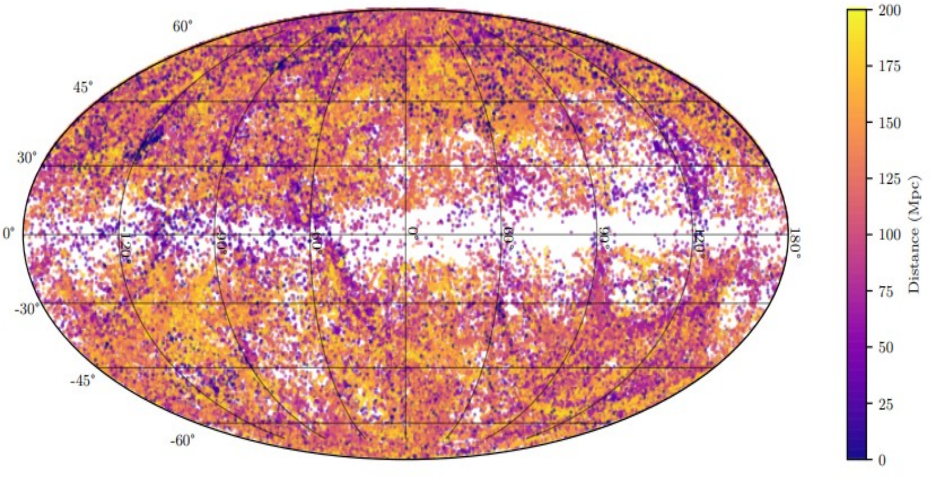


Figure 6: The distribution of the HECATE objects projected on a sky map in galactic coordinates. The different colors indicate the distances. Figure 1 of Kovlakas et al. 2021.

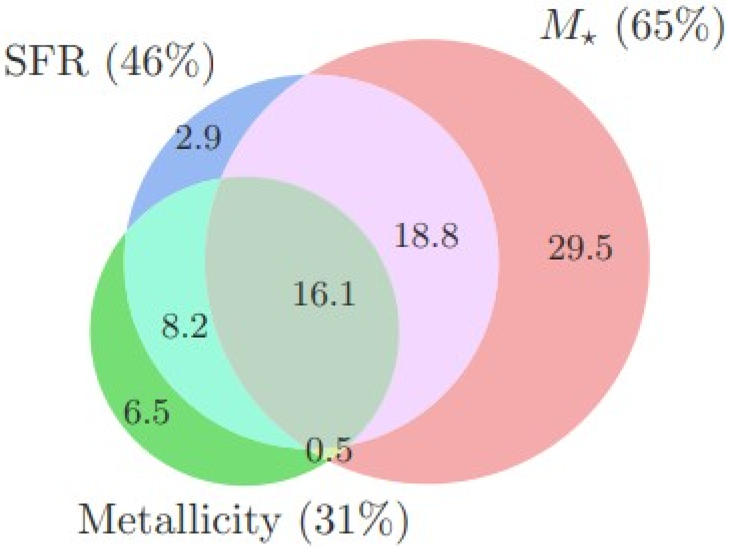


Figure 7: This Venn diagram shows the completeness of the HECATE on different stellar parameters, star-formation rate (SFR), stellar mass ($M_\star$) and metallicity. The different color areas denote the percentage of the combination of the three main stellar parameters. Figure 9 of Kovlakas et al. 2021.

The HECATE catalog contains 204733 galaxies in total. From them, 200861 galaxies will be used, as after a cross-match with AllWISE Source Catalog, some objects (3872) do not have WISE detections. Furthermore, we choose galaxies that have an activity classification (63715 of them 31.7 per cent). This catalog will be used as the basis of our analysis. The subset of galaxies with available SDSS spectroscopy (taken after a cross-match with the MPA-JHU DR8 catalog of Kauffmann et al. 2003; Brinchmann et al. 2004; Tremonti et al. 2004) will be used for the training and testing of our classification scheme, which will be applied to the overall sample. The HECATE catalog also includes WISE forced photometry for the SDSS sample.

2.3 The multidimensional classification of emission-line galaxies

For the HECATE catalog, Stampoulis et al. 2019 defined a new activity diagnostic scheme based on the location of galaxies in the 4-dimensional space (Figure 8) defined by diagnostic lines in the BPT diagrams. That scheme was based on fitting multivariate Gaussian distributions to the four emission-line ratio distributions of log([N II]/H α), log([S II]/H α), log([O I]/H α) and log([O III]/H β). The emission-line measurements were obtained from the SDSS (Sloan Digital Sky Survey; York D.G. et al. 2000). The classes of galaxies defined on that classification scheme were four: star-forming, AGN (Active Galactic Nuclei), LINER (Low-Ionization Nuclear Emission-Line Region) and composite. The criteria for the definition of each class on that four-dimensional space were defined by a Support Vector Machine (SVM) algorithm (Stampoulis et al. 2019). Based on the location of an object in this 4-dimensional space, one can determine the probability that it belongs to each of these clusters. However, in our analysis we will use their Soft Data-driven Analysis based on the class with the highest probability.

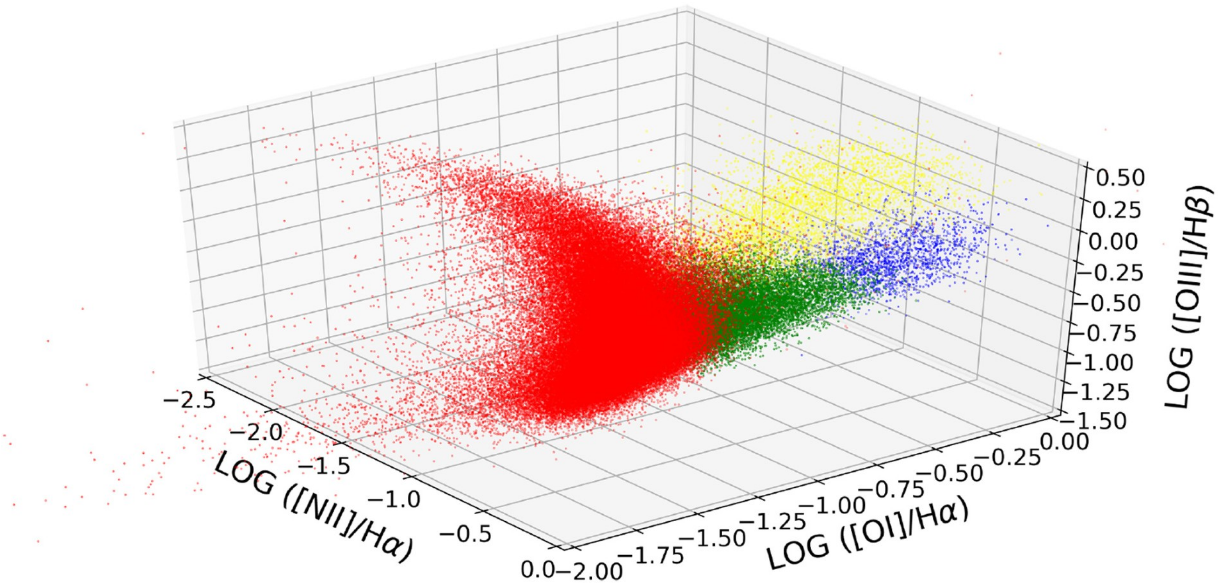


Figure 8: The three-dimensional projection of the four-dimensional emission-line space used for the multidimensional classification of galaxies by Stampoulis et al. 2019. The different classes are color-coded on that plot: red for star-forming, yellow for AGN, blue for LINERs and green for composites, figure 7 of Stampoulis et al. 2019.

The data used for that classification were taken from the SDSS DR8 data release. The SDSS is an astronomical survey in the optical light spectrum. That survey covered galaxies in 14,000 square degrees of the sky. The observations included the optical emission-lines used for the multidimensional classification of Stampoulis et al. 2019, Hα (λ6563 Å), Hβ (λ4861 Å), [OIII] (λ5010 Å), [OI] (λ6300 Å), [NII] (λ6548 Å), [SII] (λ6717 Å) and [SII] (λ6731 Å). It also performed photometry measurements for a large number of galaxies in 5 ugriz filters. The relative response of the ugriz filters is presented in Figure 9.

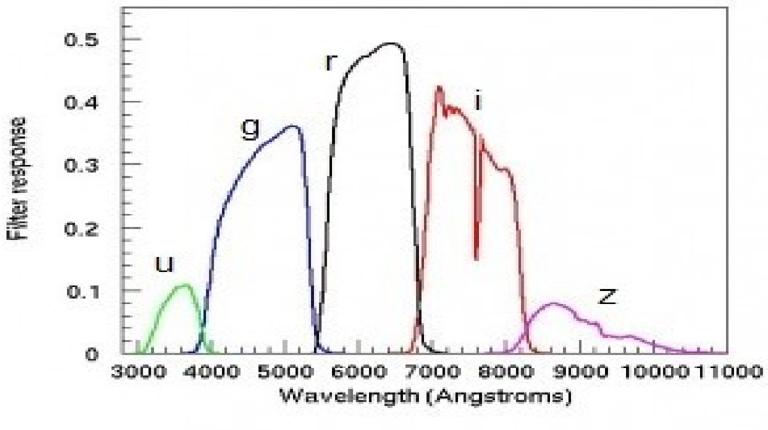


Figure 9: The relative response of the SDSS u, g, r, i and z filters. Image form the www.sdss.org.

2.4 WISE photometry scheme for the new diagnostic

Based on the Right Ascension and Declination of every HECATE object (HECATE catalog) and after a cross-match with the AllWISE Source Catalog, we obtained the w?gmag and the w?mag_2 photometry for the galaxies in the HECATE sample.

A significant fraction of our objects (about 40%) are extended sources meaning that the fix apertures of WISE will not be covering the whole angular diameter of an extended galaxy but only a portion of it. This means that for the extended galaxies (galaxies in close distances) we will record only the light emitted from their core while for point-source objects (galaxies at greater distances) we will record all the light emitted from them (Figure 10). As we have activity in both the nucleus and in the outskirts of the galaxy, this will create bias because galaxies that belong to the same class but have different distances will seem to have different emission spectrum. For this reason, it is crucial to consider a new photometry scheme that includes all the light emitted by each galaxy. One way to achieve this, is to include the w?gmag WISE photometry for those extended sources. For galaxies in further distances that will appear as point-like sources, we use the w?mag_2 fix WISE aperture to capture the light emitted from the whole galaxy.

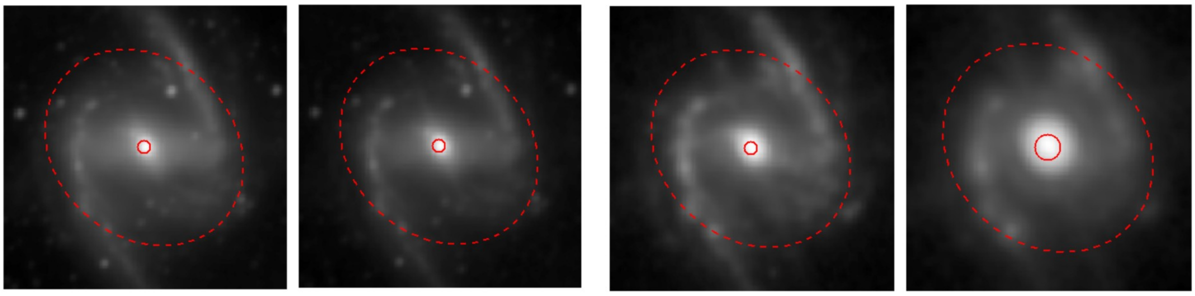


Figure 10: NGC 1365 (PGC 13179) is an example of an extended galaxy in the local Universe. The red lines, dashed and solid represent the w?gmag and w?mag_2 WISE photometry apertures respectively. From left to right the different WISE bands W1, W2, W3 and W4. If we had only considered w?mag_2 photometry aperture for our sample of galaxies, it is obvious that a lot light would be missed.

This new photometry scheme is a combination of the two WISE photometry profiles previously presented (w?mag_2 and w?gmag). For each of the four individual WISE bands, the w?gmag photometry is kept for all galaxies that have reliable photometry (based on the signal-to-noise ratio) in that aperture and the w?mag_2 photometry is introduced for all galaxies that do not have detections on the w?gmag aperture. We will refer to this photometric scheme as hybrid photometry as it combines two different WISE apertures.

The idea of a photometry that takes into consideration the whole galactic activity is not new. The "force photometry" or WF photometry (Lang et al. 2016) successfully corrects this problem for a considerable number of objects. Even though, the number of objects having their photometry corrected by this method is large, we find that in HECATE catalog, the percentage of objects that have WF photometry that we can utilize for this project (SNR>3 for the 1,2 and 3 WISE bands) are only about 40% (Figure 12). The WF photometry scheme takes advantage of the higher resolution of the SDSS survey, to locate sources on the WISE frames. In Figure 11 the one-to-one comparison (subtraction of the hybrid photometry from WF photometry in every Band) of the forced photometry scheme and the hybrid photometry scheme is presented.

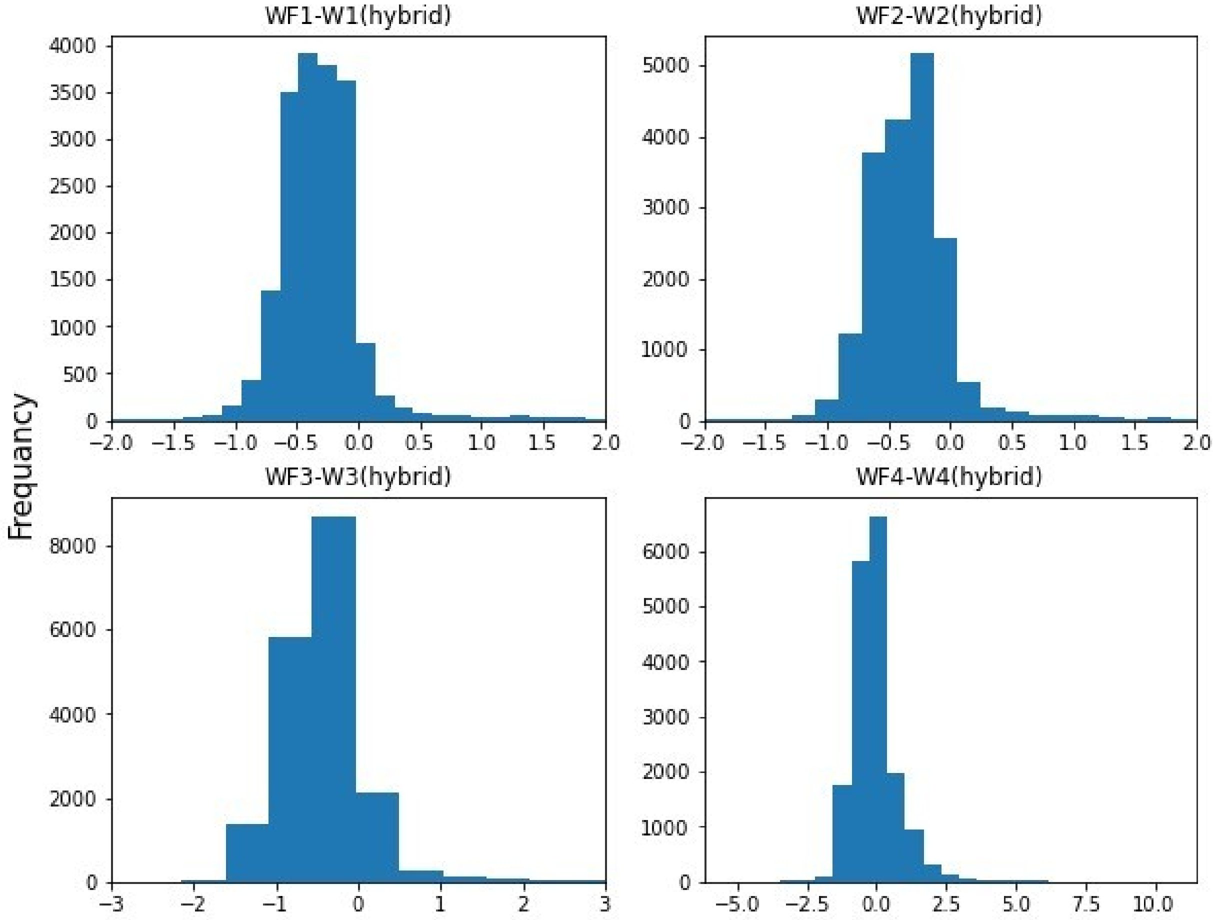


Figure 11: The distribution of the difference between the forced (WF) and hybrid (W?gmag combined with w?mag_2) photometry schemes for each WISE band. The distributions are Gaussian indicating the two photometric methods are equivalent. There is an offset about of 0.1 to 0.5 mag in the four bands.

The analytics of each WISE photometry band are presented in the plots below (Figure 12). These analytics give information about the percentage of galaxies that have detections in each band and how many of these are reliable detections based on signal-to-noise ratio (SNR). The WISE survey records the photometry in magnitudes in the Vega system. The error of each measurement is also recorded in magnitudes. The equation of magnitude is $m=-2{,}5\log_{10}(f)+c$ (1), where m is the apparent magnitude of the galaxy and f represents its flux. From this equation we can derive the signal-noise ratio through error propagation. We find that $SNR=\frac{1}{\sigma_m}$, where the σ_m is the error in apparent magnitude, for small magnitudes ~10 Vega magnitudes, as we have in this project.

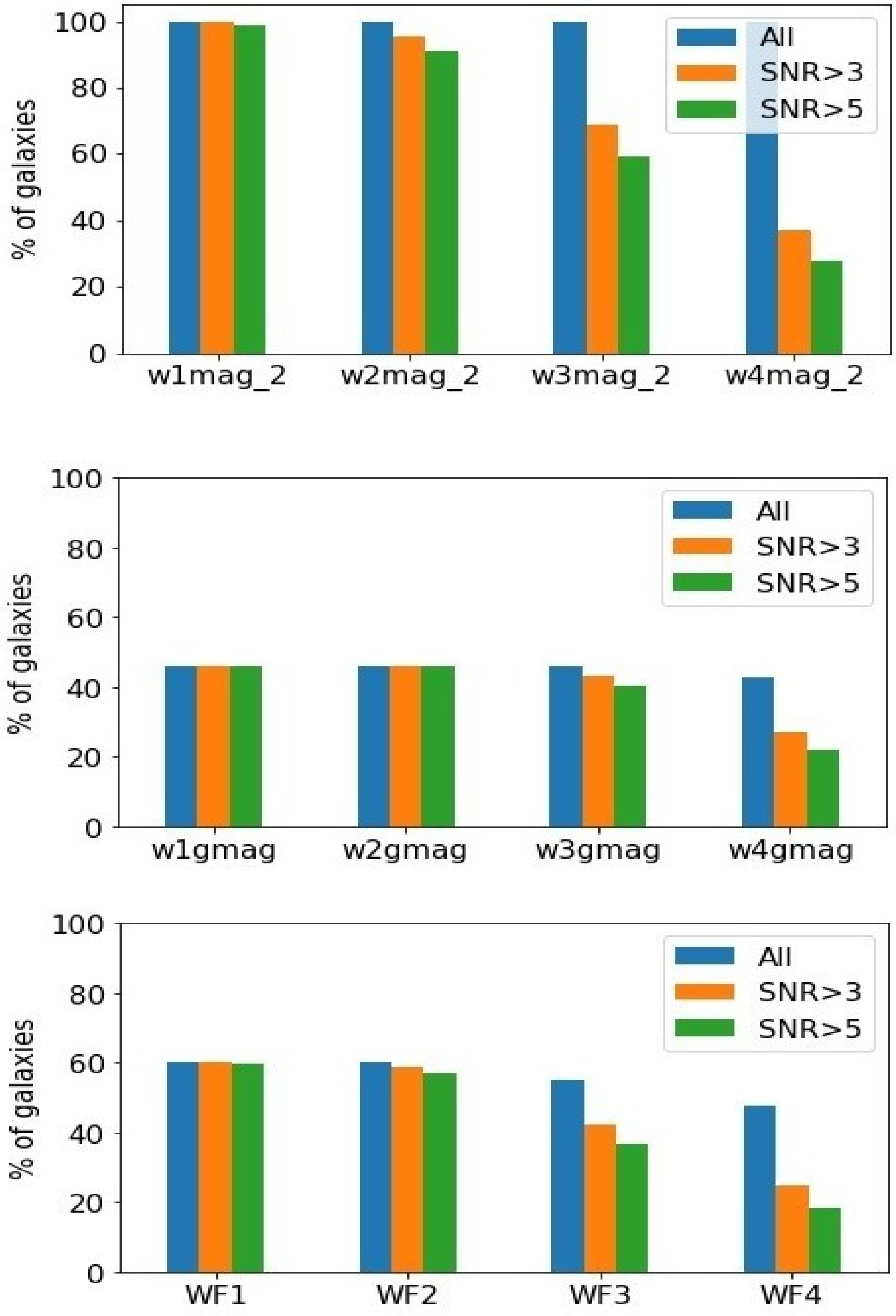


Figure 12: Percentages of galaxies in the HECATE catalog for the different photometry schemes in each WISE band: Top: w?gmag (elliptical aperture adjusted for each galaxy), Middle: w?mag_2 (fix aperture of 8.25"), Bottom WF (forced photometry method of Lang et al. 2016). Also, with different colors the percentage of galaxies with reliable photometry (orange: SNR>3 and green: SNR>5) for each band and each photometry method.

2.5 Data preprocessing

The training data for this new diagnostic are based on the galaxies that have already been characterized in different classes: star-forming, Active Galactic nuclei (AGN), Low Ionization Nuclear Emission-Line Region (LINER) and composite galaxies (section 2.3). The characterization for the first three classes was based on the optical spectroscopy from SDSS DR8 survey (see section 2.3). We consider passive galaxy any galaxy that does not show any evidence of activity. The passive galaxies were selected by the following criteria: emission-lines should have signal-to-noise ratio below 3 and the signal-to-noise ratio of the emission-line continuum above 3.

In order to properly train the new diagnostic, galaxies need to be chosen to present excellent optical spectroscopy measurements in the optical emission-lines in the SDSS survey. In that way, a high confidence level about the galaxy classification labels from the previous classification scheme can be achieved. That is important as these labels will be used for the training of the new diagnostic. For this reason, we only include the training set for the new diagnostic galaxies with signal-to-noise ratio above 5 in all the following optical emission-lines: Hα 6563 Å, Hβ 4861 Å, [OIII] (5007 Å), [OI] (6300 Å), [NII] (6548 Å) and [SII] (6717 Å and 6731 Å). Furthermore, as our diagnostic is based on the WISE bands, we need to apply an additional signal-to-noise ratio criterion for the WISE detections: signal-to-noise ratio above 5, for all classes except passive galaxies, in all three hybrid photometry WISE bands. Passive galaxies tend to be fainter and with low amounts of gas and dust their infrared emission is relatively low compared to other classes. So, for the passive galaxies the criterion for reliable photometry is the signal-to-noise above 3 in the three WISE bands. This different signal-to-noise criterion for the selection of the passive galaxies derived by the fact that a stricter criterion would drastically reduce the available sample: instead, we find that the relaxed signal-to-noise ratio >3 criterion does not. In Figure 13 we can see the galaxies that were selected as passive are in the red sequence on a color-magnitude diagram.

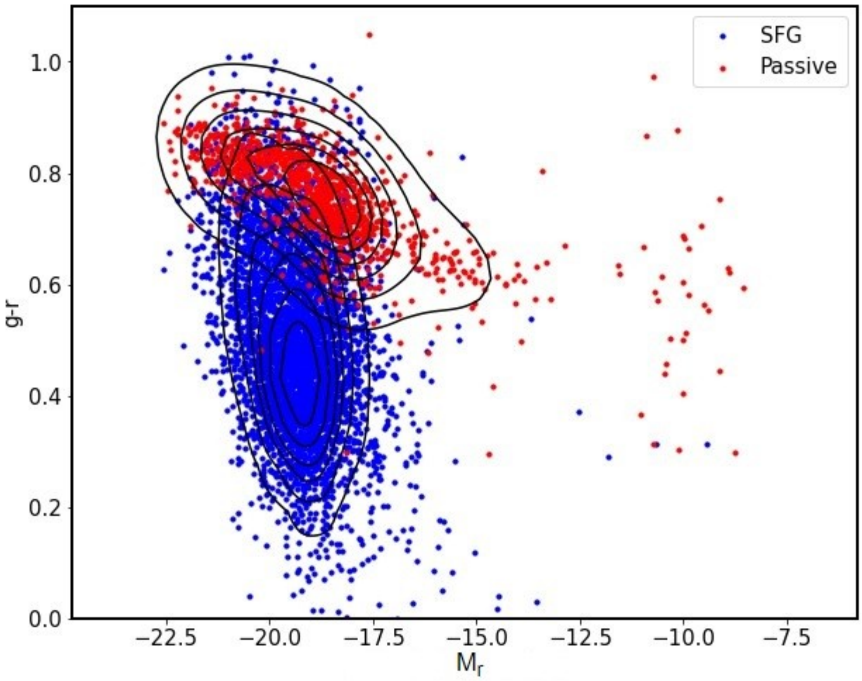


Figure 13: The color – magnitude diagram, g-r color against the r-band absolute magnitude (M_r), showing that the galaxies that was selected as passive galaxies actually belong in the red sequence. The g and r filters are referred to the SDSS survey (section 2.3).

After applying the two criteria about the signal-to-noise ratio in the optical emission-lines and the WISE hybrid photometry bands our clean data sample has 21110 galaxies in total. 15027 of them (or 71.2%) are star-forming, 1395 of them (or 6.6%) are AGN (LINERs 47% and Seyfert 53%), 1315 (or 6.2%) are composites and 3373 (or 16.0%) are passive galaxies. The merging of the Seyfert and the LINER galaxies under the category of the AGN was done under the assumption that a significant fraction of LINERs is powered by a central black hole.

Due to the nature of the passive galaxies and the fact that the WISE band 3 has relatively low sensitivity, only about a third of all passive galaxies in this sample, have reliable photometry in the W3. To expand the usable sample of passive galaxies and have an adequate number of them, we replace W2-W3 with the constant value of 99.999 for all passive galaxies that have ambiguous detections in W3 band. This replacement was done only for the passive galaxies that, even thought there was detection in WISE band 3, there also was an upper limit warning, meaning that the photometry for these galaxies in W3 was ambiguous. Also, this constant will not affect the training and the performance of the diagnostic, as the value 99.999 that was chosen (hereafter magic number), is located away from the other three feature distributions. In other words, the magic number represents non-detection in W3 band for the W2-W3 feature. Finally, the composition of the passive galaxies is the following; in total there are 3373 passive galaxies, 2465 (or 73%) have the magic number in the W2-W3 and 908 (or 27%) with normal (as measured) in all three features.

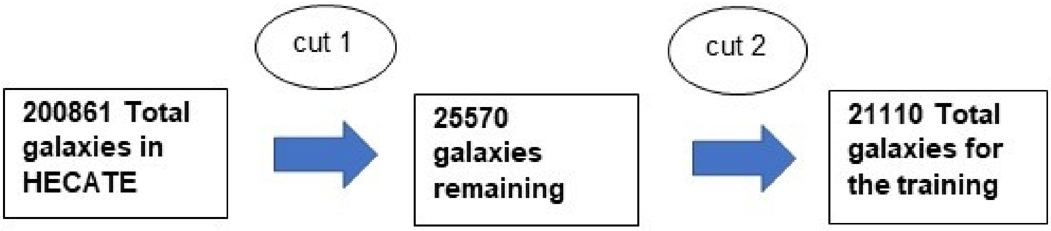

200861 Total galaxies in HECATE
cut 1
25570 galaxies remaining
cut 2
21110 Total galaxies for the training

Diagram 1: Diagram showing the process of the definition of the training sample considered in our analysis. These 21110 galaxies have SNR above 5 in all optical emission lines and in the three WISE bands 1,2 and 3 making this sample suitable for our project.

cut 1	The condition that every galaxy must present signal-to-noise ratio on the optical emission lines (Hα (λ6563 Å), Hβ (λ4860 Å), [OIII] (λ5001 Å), [OI] (λ6300 Å), [NII] (λ6548 Å), [SII] (λ6717 Å) and [SII] (λ6731 Å)) above 5
cut 2	The condition that every galaxy must present signal-to-noise ratio in the tree WISE bands above 3 for the passive galaxies and above 5 for the rest of the classes

Table 1: Selection criteria used for the definition of the training sample considered in our analysis.

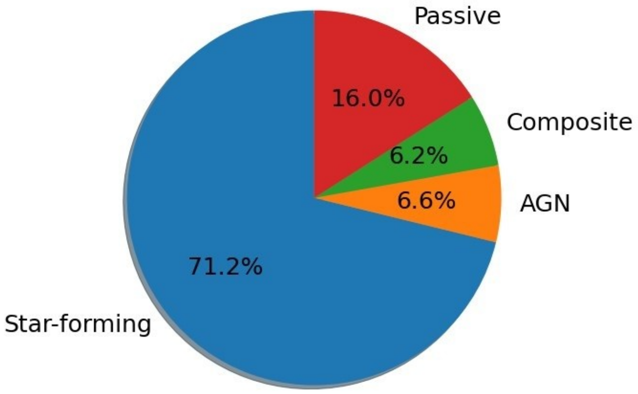


Figure 14: Training sample composition per class. It is obvious that the data sample is unbalanced.

2.6 Feature distributions

In order to choose the features that efficiently separate the 4 classes, we take into consideration the following: Assef et al. 2013 and Mateos et al. 2012 diagnostics show that color-color plots perform poorly in our sample (Figure 3 and 4). For this reason, we also introduce the luminosity information (absolute magnitude) partly motivated by the SDSS-GALEX color-magnitude diagram. We use the W2 band as a luminosity indicator instead of W3 since more galaxies have reliable W2 measurements. We find very similar results if we use W3 absolute magnitude instead of W2. If we inspect the distributions on the 2-D plots, the separation of the four classes becomes visible (Figure 15).

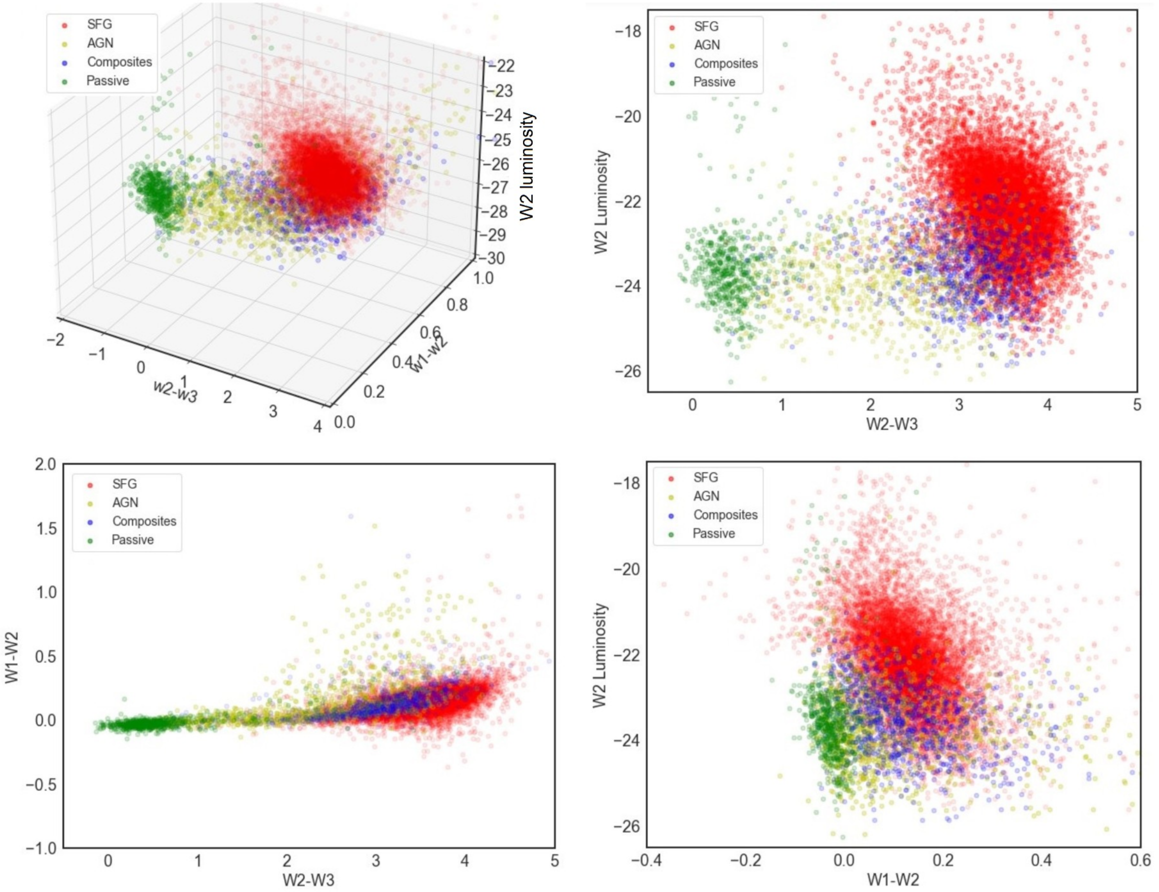


Figure 15: The feature distributions of the training data sample for the training of the algorithm. Upper left: the 3D distribution in the W2 luminosity, W2-W3 and W1-W2 feature space. Upper right: plot of W2 luminosity against W2-W3 color. Bottom left: plot of W1-W2 color against W2-W3 color. Bottom right: plot of W2 luminosity against W1-W2 color.

After choosing the features, it is important to inspect the distribution of each feature of each galaxy class, see Figure 16, in order to avoid any skew data distributions. The reason that it is important the data to be normally distributed, e.g. Gaussian, is to avoid overlapping between the feature distributions as much as possible. In this way, the Random Forest can be trained with less uncertainty allowing it to perform more efficiently on general data.

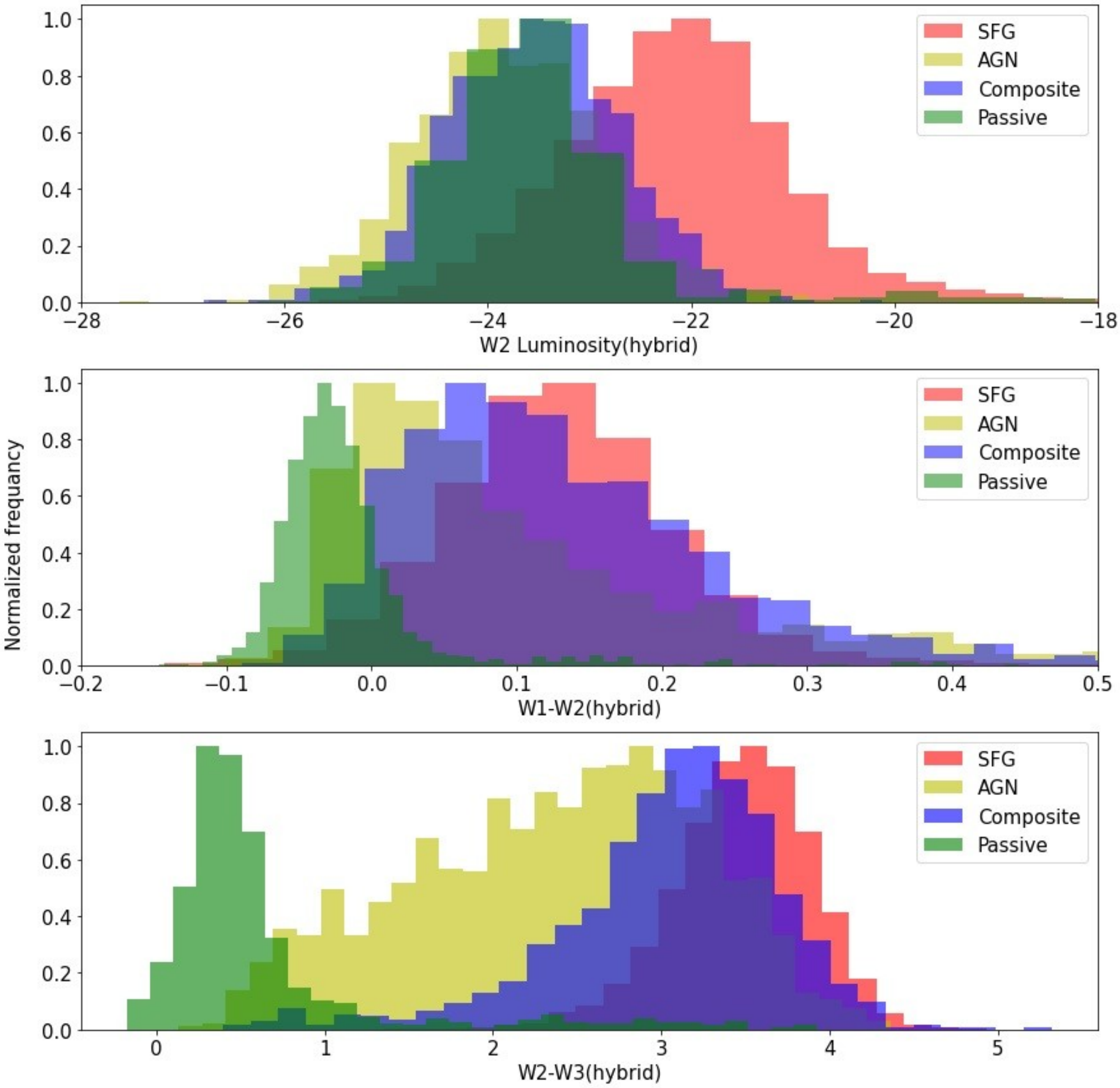


Figure 16: Distributions of the features for each class for the hybrid photometry scheme as defined for this project (section 2.4). Top: the W2 luminosity, Middle: W1-W2 color: Bottom: W2-W3 color for the star-forming (red), AGN (yellow), composite (blue) and passive galaxies (green).

3 The Random Forest classifier

3.1 The Random Forest algorithm

Although there is a variety of supervised machine learning algorithms that may be able to solve this classification problem, the Random Forest was chosen as it is more flexible than others. The limitation of an SVM is that, in this project, some classes of galaxies seem to overlap and a boundary would be insufficient to separate them. An Artificial Neural Network resembles the way the Random Forest operates but it cannot give us any insight of the underline astrophysical reasons on how the algorithm works. In the case of the Random Forest, we can inspect the "inner workings" and understand the process of its training.

In order to understand in more detail how the Random Forest algorithm works, one must start from the Decision trees. The Decision tree will start with a root node that contains all the data, then it will use the features that have been selected to progressively create more homogeneous groups of data (nodes). Ideally, at the end of the process, the final nodes (leaves) will only contain data of the same kind (class). The problem with a single Decision tree is that, in most cases, the tree tents to overfit the data. To avoid overfitting, we can combine a lot of Decision trees and build a Random Forest. Each Decision tree of the Random Forest is trained on a subsample of the data. Every subsample of the full dataset for classification that go into the trees are randomly shuffled and selected. It is called random because during the training process of the algorithm, the features are selected randomly to make the split of the data into the new nodes.

3.2 Random Forest implementation

This diagnostic uses the Random Forest algorithm. This algorithm is provided by the scikit-learn project, version 0.24.2. As mentioned before this new diagnostic is based on three features (W2 luminosity or absolute magnitude, W1-W2 color and W2-W3 color based on the hybrid photometry described in section 2.4). The algorithm discriminates between four galaxy classes, based on the three features mentioned above. The galaxy activity types we consider to be are: star-forming (SFG), Active Galactic Nuclei (AGN), composite, and passive galaxies.

The performance of the Random Forest algorithm is driven by a set of hyperparameters such as: max_depth, max_leaf_nodes, max_samples, min_samples_leaf, min_samples_split and n_estimators, see Table 2. These hyperparameters are parameters that someone can tweak and tune so that the algorithm fit different machine learning problems according to their individual needs. The values of these parameters for this implementation are based on the impact that they have on our diagnostic. Some of the hyperparameters are left on the default mode as imported from the scikit-learn package, as upon investigation, the result does not improve by tweaking them. Some hyperparameters that have significant impact in our diagnostic are the following: max_depth, max_leaf_nodes, max_samples, min_samples_leaf, min_samples_split and n_estimators. Furthermore, the value for the bootstrap hyperparameter is set to 'True', the class_weight to 'balanced_subsample',and the criterion hyperparameter to 'gini'. Due to high imbalance in the number of galaxies in each class (see Figure 14) it is important to set the class_weight to 'balanced_subsample'.

Hyperparameter	Description
n_estimators	The number of trees in the Random Forest
max_leaf_nodes	Maximum number of end nodes
max_depth	Maximum depth of every tree
max_samples	The samples chosen to train each tree
min_samples_leaf	The minimum number of objects contained in an end node (leaf)
min_samples_split	The minimum number of objects required for the splitting of an internal node
bootstrap	If true, the trees will be trained on a subsample of the original dataset. The objects are randomly selected it.
class_weight	The inverse of the frequency of appearance of the objects in each class.
criterion	Function that measures the quality of the split in each node.

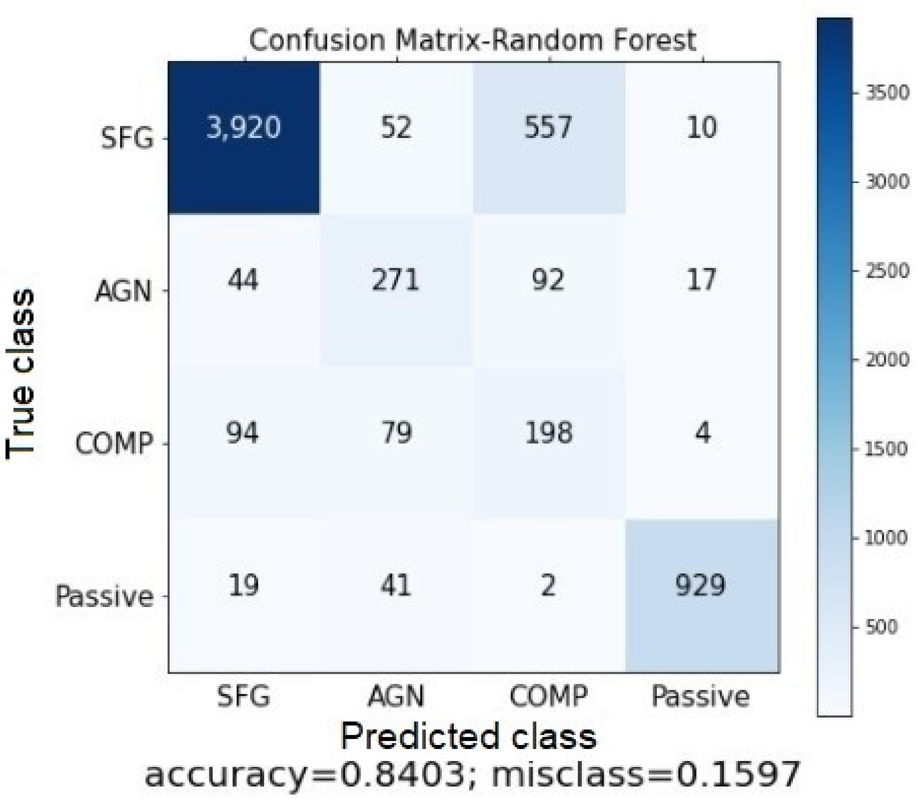
Table 2: The definition of the different Random Forest hyperparameters.

For the training process of the algorithm, the clean data sample (section 2.5) is randomly split into two data sets. The first is the training data set, which contains 70% of the original clean data sample and the other is the test data set, containing the remaining 30% of the original clean data sample. This split of the data into a training and a test set is performed on each class separately. This ensures that the training data set and the test data set will have the same composition in terms of the different galaxy classes in all four classes. The training data set is used exclusively for the training of the Random Forest. The test data set will be held out of the training process, as it is only used to check the performance data unseen by the classifier, but with similar characteristics as the parent sample. The performance metrics of the classifier and all the metrics will be calculated on that test data set. This allows us to eliminate any bias and avoid overfitting.

3.3 Classifier performance metrics

In this section, we will define the metrics used for the evaluation of the model on the test data set (similar but unseen data).

Confusion matrix: A confusion matrix gives the fraction (or number) of objects in class i that are predicted in class j after the application of the classifier. As mentioned before the classifier is trained on the training set, but the confusion matrix is calculated on the test set in order to avoid overfitting. A sample of confusion matrix for the Random Forest algorithm is presented in Figure 17.



	SFG	AGN	COMP	Passive
SFG	3,920	52	557	10
AGN	44	271	92	17
COMP	94	79	198	4
Passive	19	41	2	929

Figure 17: Example of a confusion matrix based on the application of the Random Forest algorithm on the test set of data considered in this project. An ideal confusion matrix has large values in the diagonal elements and very small values in the off-diagonal elements indicating a small fraction of miss-classifications.

In addition to the confusion matrix, we also use other metrics that give a quantitative picture of the overall performance of the algorithm.

True positive (Tp)	Items that have true class positive and have been predicted by the classifier in the same class as positive
False positive (Fp)	Items that have been mistaken (predicted by the classifier) to be positives but their true class was negative
True negative (Tn)	Items that have true class negative and have been predicted by the classifier in the same class as negative
False negative (Fn)	Items that have been mistaken (predicted by the classifier) to be negative but their true class was positive

Table 3: The terminology for the metrics used for evaluation of the new diagnostic.

For multiclass classification specifically, positive is the class of interest and negative is all the other classes. By considering positive the class under investigation and negative all the others. Table 3 contains the terminology that we need in order to define the additional metrics (e.g., Precision) for further evaluation of our diagnostic.

Accuracy is the ratio of the correct predictions to the total predictions made.

$$Accuracy=\frac{Tp+Tn}{Tp+Tn+Fp+Fn}$$

Precision is defined as the number of True positives (Tp) divided by the sum of the number of True positives and False positives (Fp).

$$Precision=\frac{Tp}{Tp+Fp}$$

The **recall** metric is defined as the number of True positives (Tp) divided by the sum of the number of True positives and False negatives (Fn).

$$Recall=\frac{Tp}{Tp+Fn}$$

The **f1-score** is defined as the harmonic mean of the precision and recall.

$$f1-score=\frac{2*Precision*Recall}{Precision+Recall}$$

The **False Negative Rate (FNR)**, is the fraction of the true positive examples that are classified incorrectly:

$$FNR = \frac{Fn}{Tp + Fn}$$

Specificity is the measure of how often the classifier predicts positive result while the true condition is positive:

$$Specificity = \frac{Tn}{Tn + Fp}$$

False Positive Rate (FPR) is the amount of the negative predictions that the classifier predicts when the true label is negative:

$$FPR = \frac{Fp}{Tn + Fp}$$

3.4 Classifier optimization

In the previous section it was mentioned that only a few of the Random Forest hyperparameters significantly affect its performance. The hyperparameters that have real impact and hence are worth optimizing are the following: max_depth, max_leaf_nodes, max_samples, min_samples_leaf, min_samples_split and n_estimators. The process of optimization is performed with the use of the GridSearchCV algorithm which is also provided by the scikit-learn Python package. The determination of the optimal values for the hyperparameters is usually done by training the algorithm several times with different choice of the hyperparameter values each time, typically by means of a grid search. The performance of the algorithm on the test set is evaluated in each point of the grid, and the optimal set of parameters that maximizes the performance is chosen. However, since a broad grid search in a 6-dimensional space can be very computationally intensive we first narrow the range of the parameters by calculating the performance for different values of each hyper-parameter separately. This way we calculate the validation curves which show different performance metrics as a function of each hyperparameter values. Once we have determined the sensitivity of the algorithm on the different hyperparameters and the ranges of the parameters that significantly affect its performance, we perform a grid search around these ranges. The range for each hyperparameter (Table 4) was found by inspecting the behavior of the two scores, on the training set and in the cross-validation. Where the two scores diverge means that the hyperparameter have optimal value, while when the two scores diverge the algorithm starts overfitting the data.

For the evaluation of the performance, three crucial metrics, the f1-score, the precision and the recall (defined on section 3.3) for each galaxy class and each hyperparameter are monitored. Plots of these validation curves, for each minority class (all classes except the star-forming) are considered to validate that there is no overfit of the data (Appendix A). The reason for plotting these validation curves, beyond the purposes of saving computational time and resources, is that we have a complete supervision on how the grid search should perform.

More specifically, each plot has the performance score of a metric (e.g. f1-score) on y-axis and a range of the possible values of one hyperparameter (e.g. n_estimators) the x-axis while keeping all the other hyperparameters constant. The performance reported on each plot is the performance on the training data and the other is the performance on a data set with cross-validation (CV). It is obvious from the plot that the performance on the training data is misleading, as the algorithm tents to overfit as it adapts patterns to recognize only these data, allowing it to achieve perfect accuracy for that particular data, but poor performance to similar unknown data. The best hyperparameters are found in the areas of the plots were the two lines (training set score and cross-validation score) are converging. Large distance between them means overfitting. To determine the ranges of the hyperparameters that the grid search will explore to find the best values, the plots of model accuracy against a wide range of hyperparameters are considered (Figure 18).

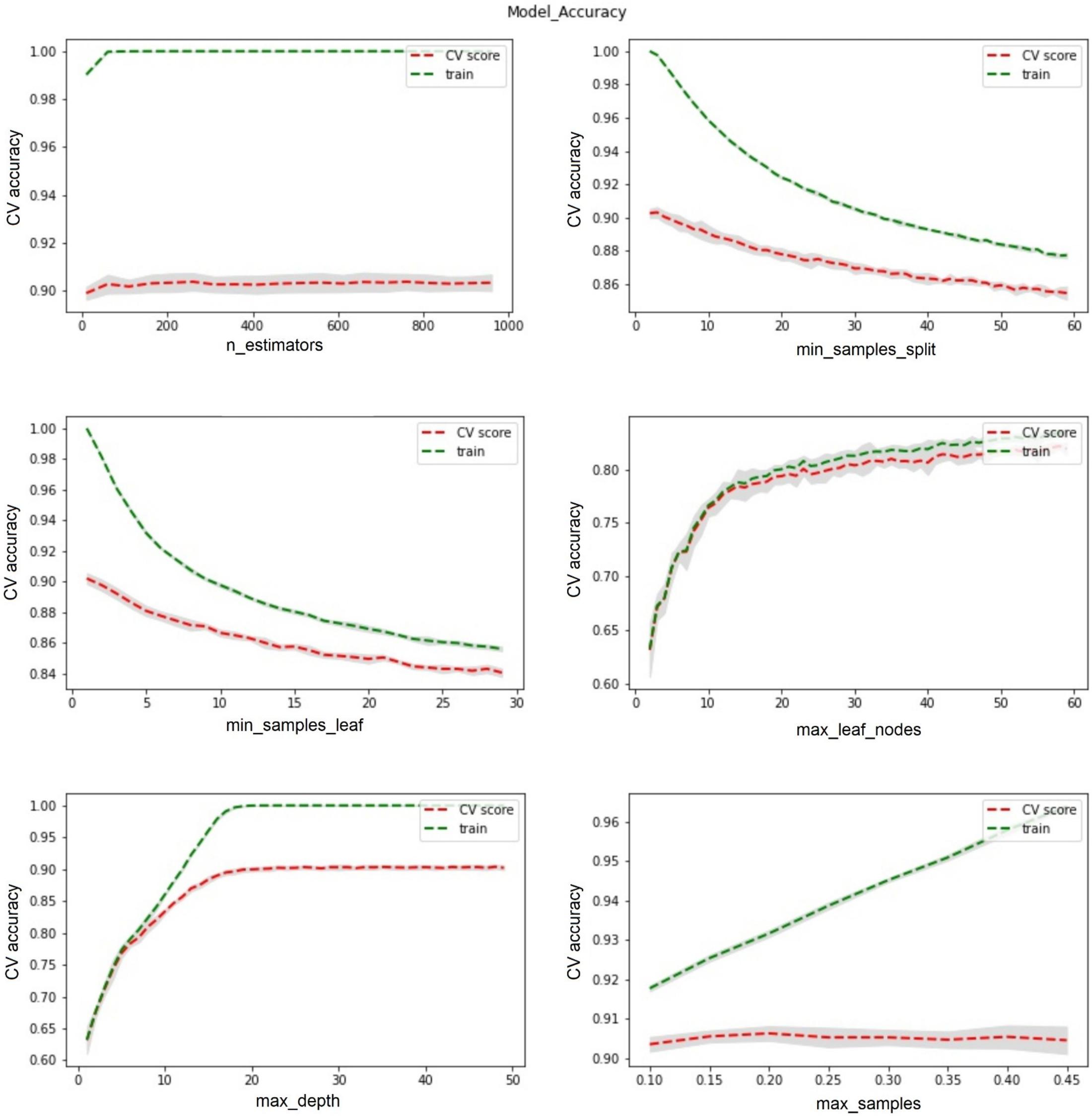


Figure 18: Validation curves for Random Forest classification accuracy as a function of each of the six hyperparameters. Each plot represents a different hyperparameter. The red dashed line represents the score on the training data set and the green dashed line with the 5-fold cross-validation method. The shaded area represents the uncertainty (standard deviation) on the 5-fold cross-validation.

The next step is to use these best value ranges for each hyperparameter extracted by the plots and with the grid search to find the optimal combination of them. The optimal values are presented in Table 4:

Hyperparameter	Grid search range	Best Value
n_estimators	10 - 300	250
max_leaf_nodes	20 - 40	34
max_depth	10 - 15	13
max_samples	0.01 - 0.2	0.1
min_samples_leaf	10 - 20	16
min_samples_split	25 - 40	30
bootstrap	-	True
class_weight	-	balanced_subsample
criterion	-	gini

Table 4: The best values for the nine hyperparameters of the Random Forest, six as derived by the grid search algorithm, including the ranges where the search was performed.

One check of the stability of the algorithm, is to perform a K-fold cross validation and observe the change of the accuracy score. This is possible by splitting the data into K parts (folds), each one consisting on 4222 objects (for K=5), fit the Random Forest to the K-1 folds (training) and check the performance on the K fold (test set). We repeat this process until each one of the individual folds has have been in the position of the testing fold. Each time, the accuracy score is recorded. At the end, the average and standard deviation of these accuracy scores is calculated. A stable algorithm should have low standard deviation on its accuracy which means that the performance of the algorithm does not significantly fluctuate between its subsequent application on similar data, although small fluctuations are unavoidable as a result of the stochastic nature of the algorithm. Even though the sample is fairly large, the data for some minority classes are low. For that reason, the number of folds chosen here is 5. The average value and standard deviation for the accuracy score on the K-fold cross-validation for this project is 84.2±0.04%.

Another way of evaluating the performance of the model is the plot of Receiver Operating Characteristic or ROC curve. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity). The area under the curve (AUC) is the ability of the algorithm to separate between the classes in a binary classification (Figure 19). The higher the AUC value for a class the better the ability of the algorithm to distinguish that class from the other. Because this diagnostic has four classes the method for plotting the

ROC is to use the method one-against-all. In other words, the sensitivity and specificity are calculated for each class separately as it would be a binary classification, meaning that the class under investigation is one class and the rest of them are grouped together to form the other class.

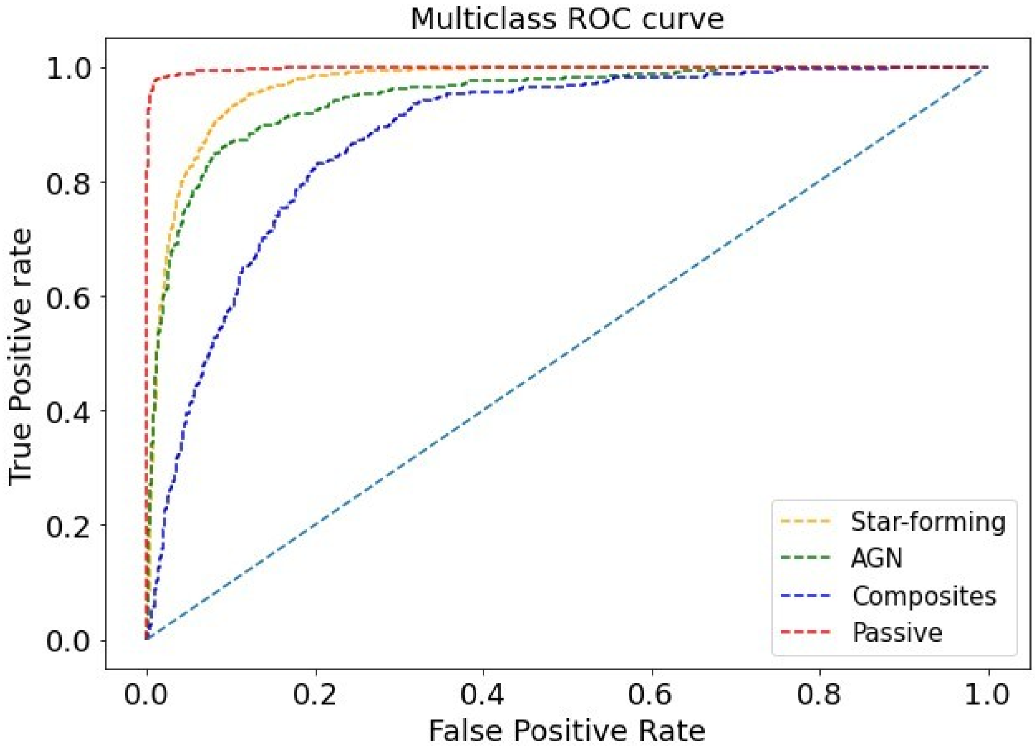


Figure 19: Example of a ROC curve. The plot of the True Positive Rate against the False Positive Rate. As expected, the new diagnostic is capable of distinguishing passive and star-forming galaxies more efficiently than the other classes. The light blue line represents a line of slope 1 or AUC=0.5, indicating a classifier that predicts randomly.

3.5 Feature importance

The feature importance of the Random Forest describes the relevance of each feature for the training of the classifier. The relevance is the measure of how much a specific feature contributes to the ability of the Random Forest to discriminate between the different classes. So, a feature that easily separates the classes will have high relevance (or importance). It is vital to check the importance of the Random Forest features as by using the smallest amount of them, facilitates better applicability of the classifier in a wider range of datasets. Furthermore, it enables the determination of the physical parameters that drive the performance of the classifier which can lead in the design of more efficient classifiers. In addition, including more features with relatively low importance score can result in poor generalization of the model. This will result in poor performance on data that are similar but different with the ones that it was trained on.

4 Discussion of the results

4.1 Performance of the classifier

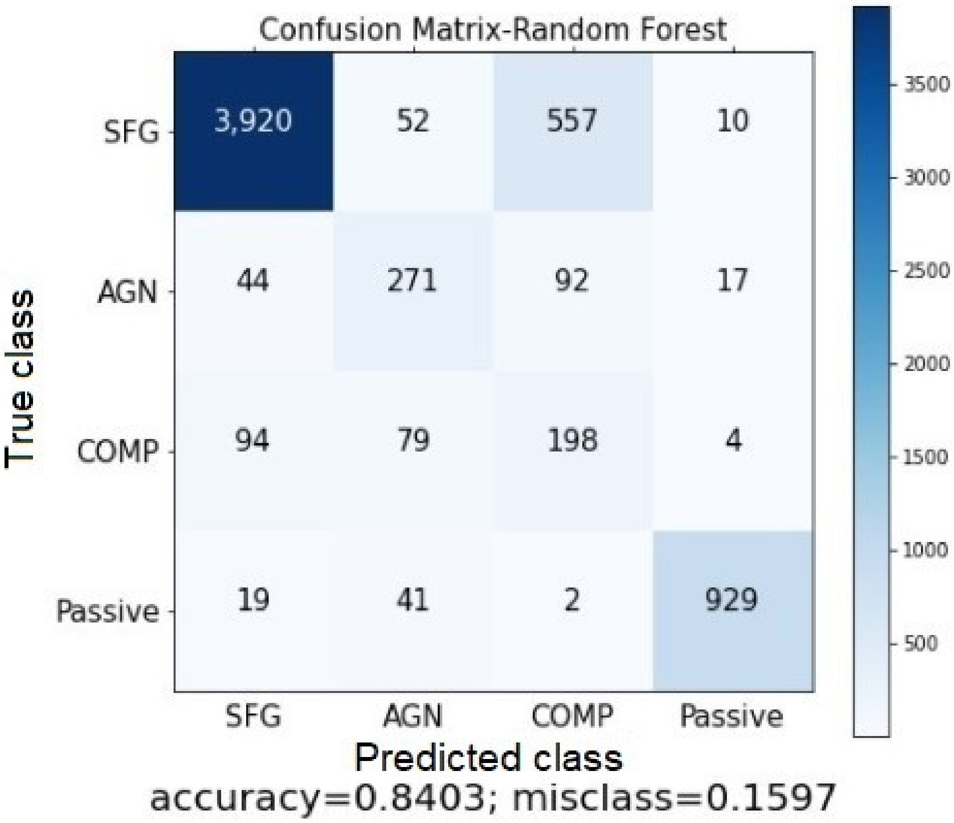
We present the results of the performance of the algorithm on our clean data sample. The algorithm is trained on a training set and tested on a test set determined based on a 70% - 30% split of the original sample (section 2.5). We used the optimal values of the hyperparameters described in sections 3.3 and 3.4. The overall accuracy we achieve is 84%. We summarize the performance metrics of the Random algorithm in the Table 5.

	Precision	Recall	F1-score
star-forming	0.96	0.86	0.91
AGN	0.61	0.64	0.63
composite	0.23	0.53	0.32
passive	0.97	0.94	0.95

Table 5: Performance metrics for the 4-class Random Forest on the test set.

From Table 5 we can see that the performance of the algorithm for the star-forming and passive galaxies is excellent. As for the other two classes of galaxies, AGN and composite galaxies, the scores are descent. These results can be explained by the fact that the star-forming and passive galaxies, have distributions that are far from the other classes and also with low scattering (section 2.6). On the other hand, the extensive mixing between AGN and composite in the feature space, as seen in Figure 15, creates confusion for the algorithm.

The confusion matrix (Figure 20) was calculated on the test data set and is a good measure of performance on general-unknown data as these was never part of training process of the Random Forest. The SFG (star-forming galaxies) and the passive galaxies have low misclassification instances. Some of the problems observed in this confusion matrix is the fact that some star-forming galaxies are classified as composites. There is also a mixing between the AGN and composite galaxies.



	SFG	AGN	COMP	Passive
SFG	3,920	52	557	10
AGN	44	271	92	17
COMP	94	79	198	4
Passive	19	41	2	929

accuracy=0.8403; misclass=0.1597

Figure 20: Confusion matrix based on the application of the Random Forest algorithm on the test data considered in this project.

The ROC curve (Figure 21) for the algorithm presented in this project (section 3.4) can give information on how well our diagnostic can distinguish every class. We find that our diagnostic has excellent ability to discriminate star-forming and passive galaxies among the other galaxy classes in the sample.

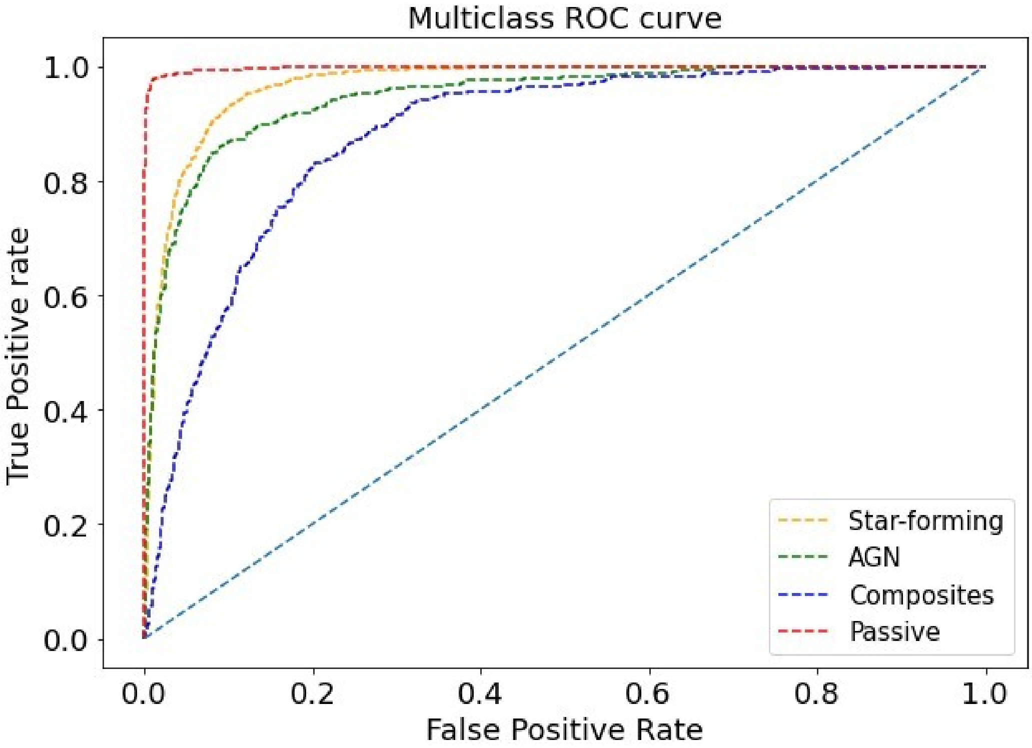


Figure 21: ROC curve for the diagnostic based on the Random Forest algorithm for this project.

Another measure of performance could be the comparison between the percentage composition per class considering the true labels in the test sample, with the composition of the same sample but with the prediction for each galaxy made by the new diagnostic. In other words, on the test sample, we must find the same composition per class when calculating it based on the true labels or with the ones predicted by the new diagnostic. In section 2.5; Figure 14, we found the composition of the test sample based on the original labels. The test set is a part of the training sample that was not used in the training process. This is the sample that we kept separate to evaluate our model in data similar to the training but the algorithm was never trained on. The composition of the test set, per class, is the following: star-forming 71%, AGN 7%, composites 6% and Passive 16%. For the same sample of galaxies, the Random Forest predicted the class for every galaxy finding that the percentage composition per class is star-forming 65%, AGN 8%, composites 12%, and passive 15%. In Table 6, we see that the composition of the test sample is similar for the passive and AGN galaxies. There is also a migration of a small population of star-forming to composites.

CLASS	Test sample original composition (%)	Test sample Random Forest composition (%)
star-forming	71	65
AGN	7	8
composite	6	12
passive	16	15

Table 6: The comparison of the test sample composition per class as it was originally in our test set (before the training) and the composition per class of the same sample of galaxies but with the classification performed by the Random Forest.

4.2 Probability distributions

Besides the classification of each galaxy, the Random Forest algorithm can also give an estimation of the probability of an object belonging to each one of the classes individually. As we know the Random Forest consist of many decision trees. For the galaxies in the sample, each tree, takes as input a galaxy and gives as output (or vote) the class that this individual galaxy belongs. Then, this process continues until that galaxy has been through every tree of the ensemble. In the end, the decisions made by every tree of the ensemble for the galaxy under question are summed. That galaxy belongs to the class that collected the most votes after the decisions (votes) of the trees. The algorithm also allows us to calculate the probability of that object to belong to each one of the classes. This probability is given by the ratio of the number of votes the galaxy received to belong in a particular class to the total number of trees considered in the algorithm.

In order to evaluate the confidence of the classifications performed by the diagnostic, we compare the probability of the highest and the second-highest ranking class for each galaxy (see e.g. Stampoulis et al. 2019). Objects with high probability in one class and relatively low probabilities in the rest of them, is an indicator of a confident and reliable classification. These plots are shown in Figures 22 and 23 for all objects and for every class individually. Objects appearing in the top right corner of that diagram have reliable classification as they have high probability to belong to a class (close to 1) and the difference from the second highest probability is low. It is calculated that for the galaxies in the test set, 51.7% of the objects have maximum probability above 75%.

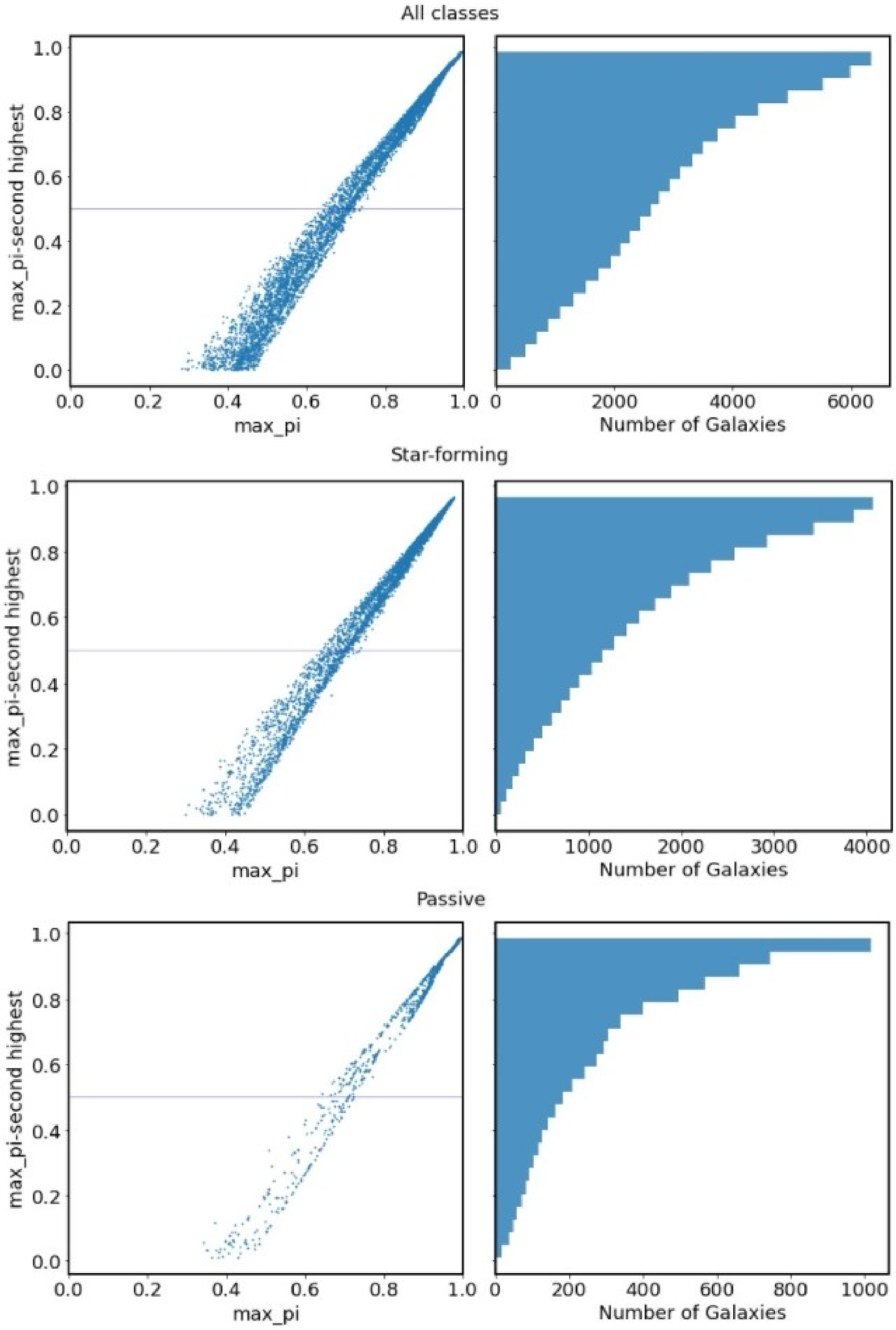


Figure 22: All plots located on the left, are plots of the maximum minus second largest probability of an object to belong to a class against the maximum probability (max_pi). The blue horizontal line represents difference in maximum probability minus the second maximum probability of 50%. All plots on the right, are cumulative plots of the same as the ones on left. According to the two plots in the top row that refer to all objects we find that 85.8% of all objects present maximum probability above 50%, and 51.7% above 75%. In the other two plots in the middle row that refer to star-forming

class indicate that 92.7% of star-forming galaxies have maximum prediction probability (as predicted by the Random Forest to belong in a particular class) above 50%, and 63.7% above the 75%. For the passive galaxies in the bottom row we have that 95.1% of them have maximum prediction probability above 50%, and 75.5% above 75%.

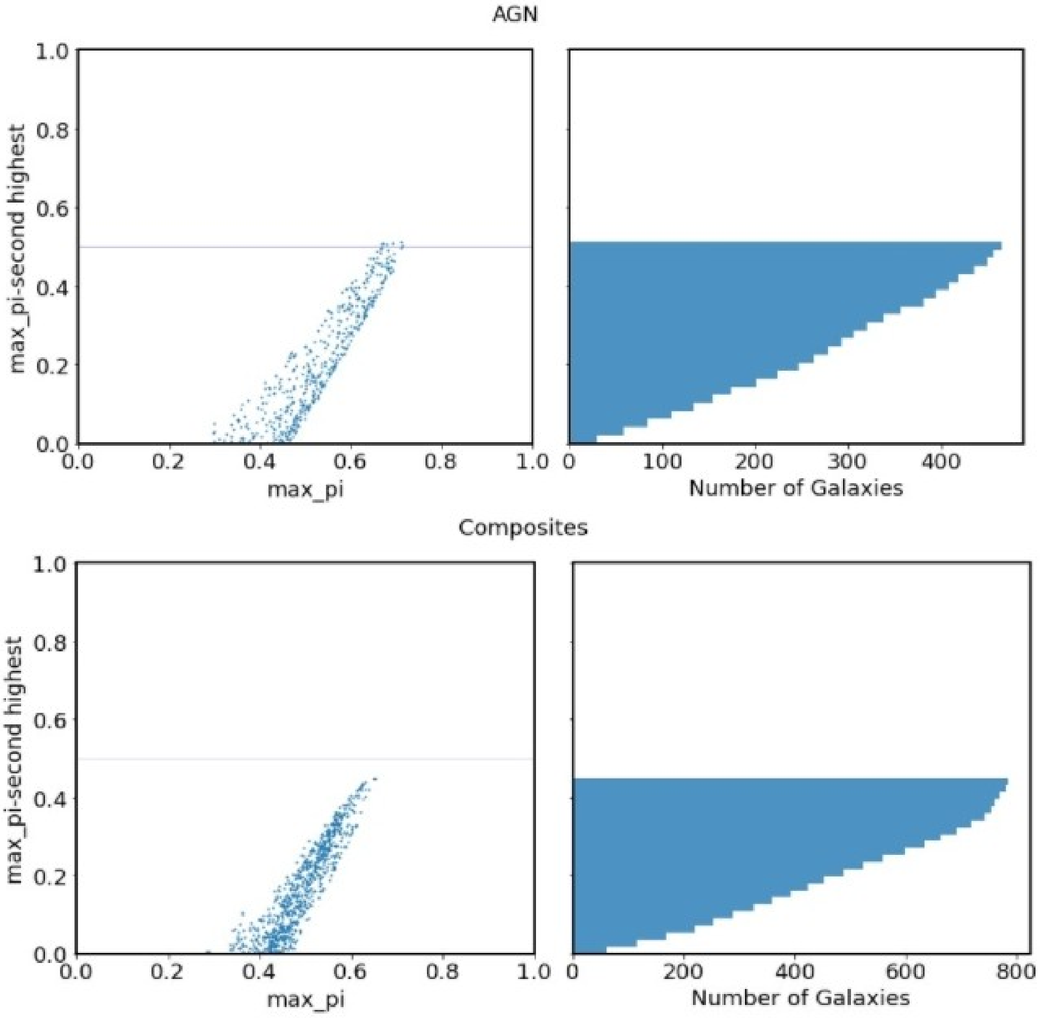


Figure 23: Probability distributions for Random Forest predictions of the class of AGN and composites galaxies. Similar to the ones above (Figure 22). The two upper plots (top row) refer to the class of AGN while the two bottom ones refer to the class of composite. From the galaxies that predicted as AGN by the new diagnostic, 62.2% of them have maximum prediction probability above 50%, and none above 75%. For the galaxies predicted as composite, 45.0% of them have maximum probability above 50%, and none above 75%.

Despite the success of the algorithm to predict accurately and with high confidence star-forming and passive galaxies, it is clear that there is some uncertainty for the classes of AGN and composites. For that reason, we also provide an alternative diagnostic (see Appendix B) that only considers three classes: star-forming, passive and other emission-line objects or “other”. In that “other” category, the classes of AGN, LINERs and composites have been merged. The motivation behind this is the observation of the extensive mixing of these classes in the 3-Dimensional feature space (see Figure 15).

4.3 Reason of the success

The reason why this new diagnostic works better than the previous ones is the amount of information that includes. In the figure 24 we see that the four classes of objects present different behavior in the region of the infrared spectrum covered by the WISE bands. Furthermore, if we examine the feature distribution of W1-W2 (Figure 16) we observe that, for the passive galaxies, the W1-W2 color has a peak around –0.4 which seems to agree with the SED (Spectral Energy Distribution) of the upper right plot of Figure 24.

In particular, we see significant emission from Polycyclic Aromatic Hydrocarbons (PAHs) lines in the W3 band in the case of star-forming and composite galaxies. On the other hand, passive galaxies are poor in dust and the populations of stars are old, resulting in declining emission in redder wavelengths. In contrast, AGN show rising emission in the mid-infrared in all WISE bands. That can be explained by emission from the accretion disk that is reprocessed by the circumnuclear dust present around the black hole. The accreted material heats up as it falls inwards to the Black Hole. For an AGN the major contribution in the IR spectrum (IR bump) is by the dusty torus around the accretion disk of the black hole while they have weak PAH emission since these sensitive molecules are destroyed by the strong UV radiation from the accretion disk (e.g. Alonso-Herrero et al. 2014) or their emission is diluted by the AGN continuum (e.g. Genzel 1998).

Composite galaxies have weaker continua in the 3-12μm range than AGN, but with stronger PAH emission, which however is weaker than that of star-forming galaxies. They also show strong silicate absorption (at ~10μm). This is reflected in their W1-W2 and W2-W3 colors which are intermediate to those of AGN and star-forming galaxies (Figure 24). From Figure 24 we also see that the W2 luminosity, is able to separate the classes of AGN and star-forming as the former will generally have higher fluxes in the W2 WISE band.

Passive galaxies also have large luminosities, but their bluer W1-W2 and W2-W3 colors discriminate them from the similarly luminous AGN. Composites have intermediate luminosity to AGN and star-forming. Also, they lie between AGN and star-forming in all considered features, resulting in their weaker performance in comparison to the other classes. Finally, since in our diagnostic we use the integrated emission of the galaxies (in order to avoid aperture effects) its application is not limited to the local Universe.

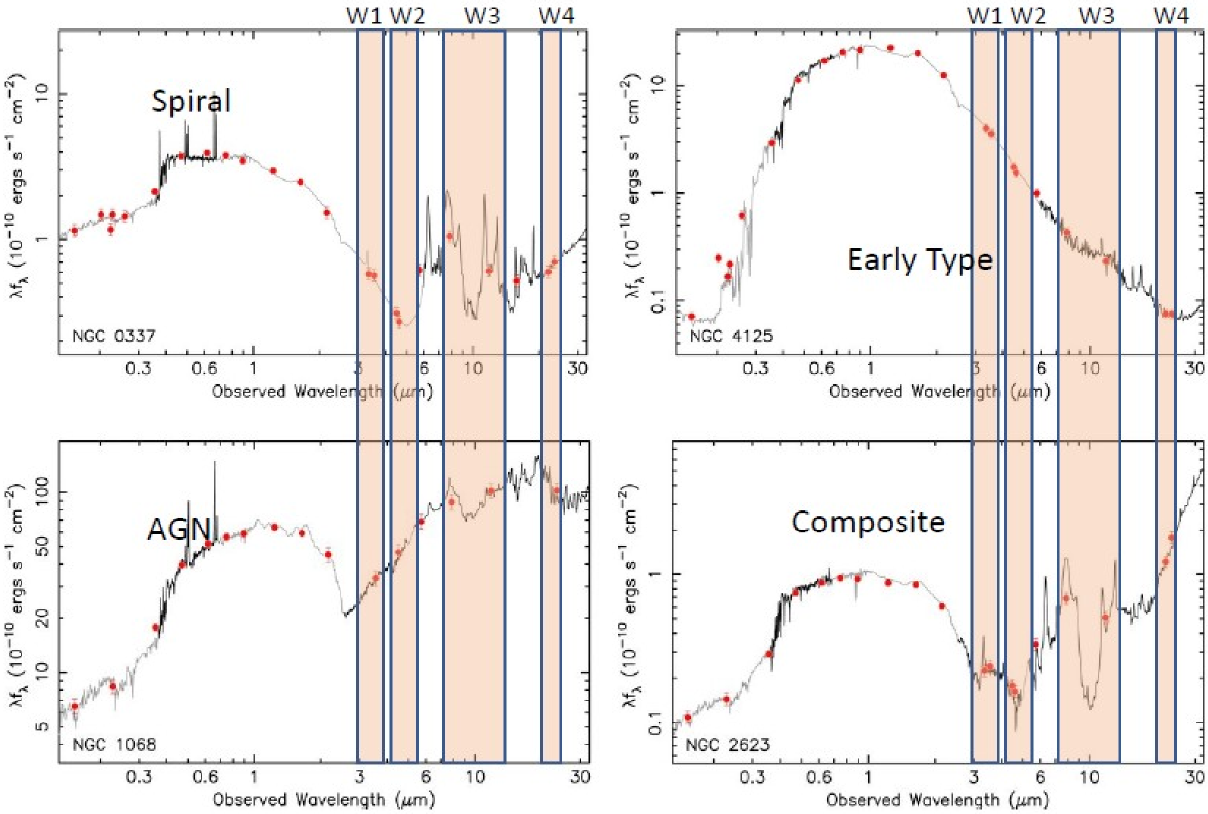


Figure 24: Brown et el. 2014. SEDs (Spectral Energy Distribution) for each galaxy class. The different WISE bands are indicated in the shaded areas on the plots.

In the previous analysis (section 4.1) we saw that the Random Forest diagnostic managed to achieve an overall accuracy of 84%. That means that the new diagnostic classified 84% of the galaxies correctly. In Figure 25 we can see that despite the success of the diagnostic it still is in disagreement with the W1-W2 > 0.8 criterion of Assef et al. 2013 as well as with the AGN locus of Mateos et al. 2012.

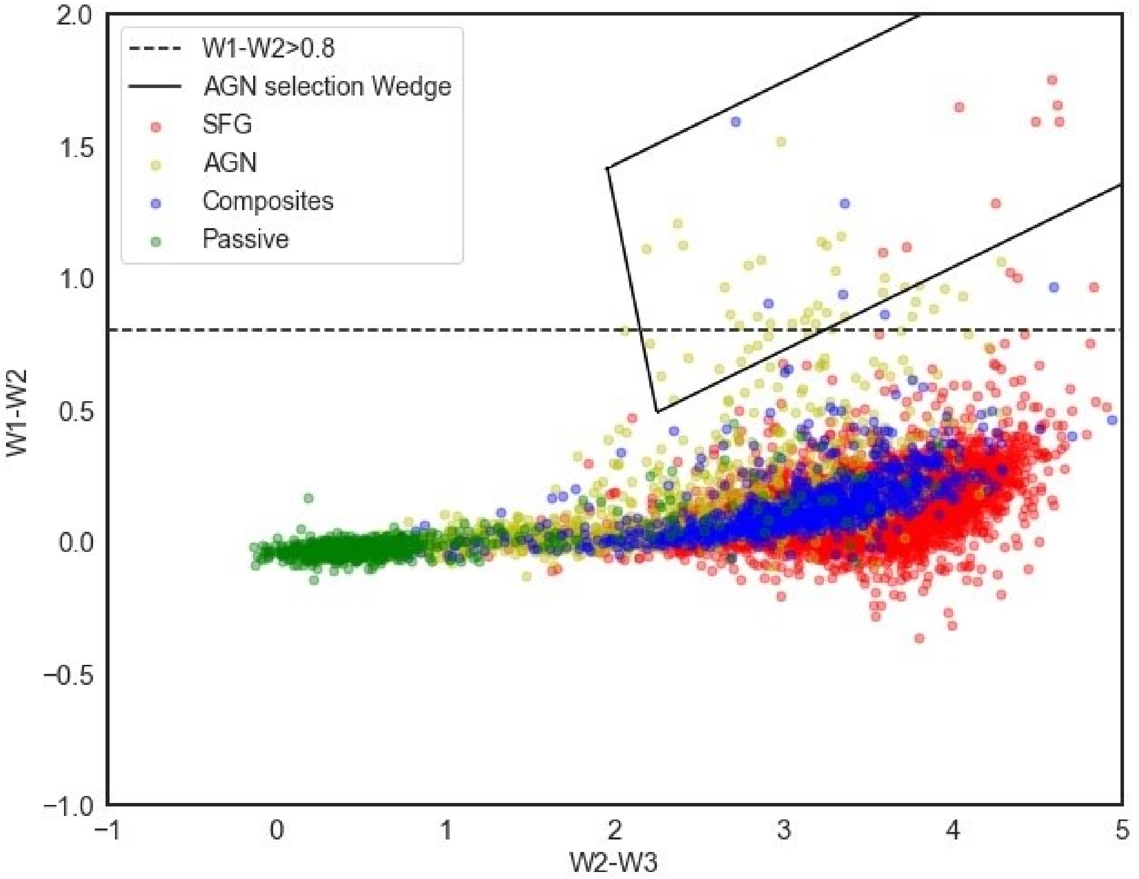


Figure 25: Plot of W1-W2 color against W2-W3 color for the training data (SDSS subsample) used in this diagnostic. The W1-W2 > 0.8 criterion of Assef et al. 2013 as well as with the AGN locus of Mateos et al. 2012 is visible on the plot. The classification of these objects was performed with the new diagnostic.

Now, considering everything mentioned above, we can check the feature importance (section 3.5) to see what features were important during the training of the algorithm. Features that have high importance will have higher score as these ones helped the Random Forest to discriminate the different classes more efficiently. From the Figure 26 below, it is obvious that the most important feature is the W2-W3 WISE color, followed by the absolute magnitude of W2 WISE band and W1-W2 WISE color.

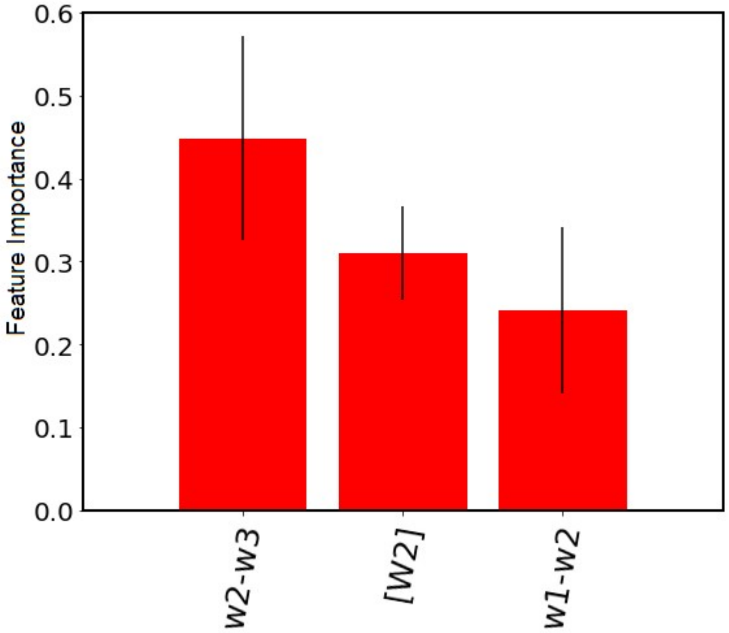


Figure 26: Feature importance. The red bars represents the value of importance of each feature and each black line on top of red bars is the standard deviation.

Observing the Figure 26 seems that the feature which has the highest impact is the W2-W3. This can be explained as the star-forming galaxies have a lot of newly formed stars which heat up the abundant dust clouds found in them. The surrounding dust heats up by UV radiation emitted by the new stars. As a consequence, these dust clouds cool by emitting in mid-infrared spectrum, and in particular, in W3 WISE band as the dust grains and PAHs have infrared emission bands lines at the ranges of ~3.3μm to 10μm. That strong emission in the W3 band makes the discrimination between the class of star-forming galaxies from the rest of them clearer for the Random Forest.

4.4 Systematic behavior as a function of sSFR

In order to obtain a better insight into the performance of the algorithm we explore its behavior as a function of different physical parameters of galaxies. In these comparisons we use as reference the classification of galaxies on the basis of their nuclear activity using the multi-dimensional optical line-ratio diagnostics of Stampoulis et al. 2019 (which are also used for the determination of the labels used in our training and test samples). We compare the fraction of star-forming galaxies that our diagnostic classifies as star-forming, composite, AGN and passive galaxies as a function of star-formation rate and specific star-formation rate (sSFR). The latter are obtained from the work of Salim et al. 2016. The sSFR is defined as the star-formation rate of a galaxy (the mass of gas turned into stars and is measured in $M_{\odot}$/yr) divided by its total stellar mass. The motivation behind this comparison is that the fraction of correctly classified star-forming galaxies is expected to increase with increasing sSFR, as star-forming galaxies are expected to have a relatively higher sSFR than every other classes of galaxies presented in this project.

Figure 27 shows the fraction of star-forming galaxies (based on the optical line-ratio classification) that we classify in each of the 4 classes we consider, as a function of their sSFR. We adaptively group the galaxies in bins of sSFR with each bin containing at least 400 galaxies. For each bin we calculate the fraction of star-forming galaxies that are classified as star-forming, AGN, composite, or passive galaxies.

We find that for galaxies with higher sSFR the classification of star-forming galaxies is very reliable with 100 per cent recall rate. On the other hand, as the specific star formation rate drops, the diagnostic systematically classifies these as composite instead of star-forming. Unsurprisingly, the fraction of star-forming galaxies classified as passive is effectively 0 while the fraction of star-forming galaxies classified as AGN rises in the lowest sSFR bins, but it is dominated by significant uncertainty.

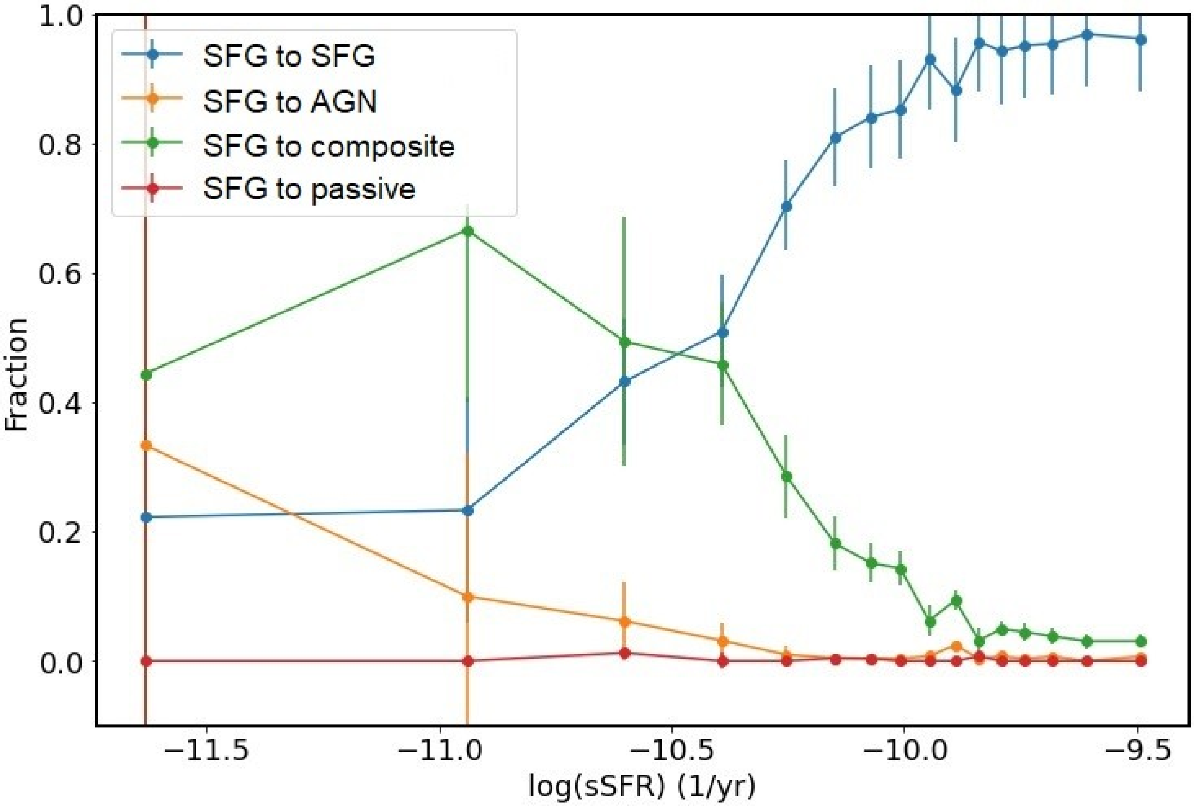


Figure 27: Plot of the fraction of galaxies as a function of specific star formation rate log(sSFR). The fraction is calculated by dividing the number of galaxies that the diagnostic predicted to be in a particular class by the number of true star-forming galaxies (as defined by Stampoulis et al. 2019) as function of log(sSFR) in every bin. Each bin contains about 400 galaxies. Every point of the plot represents the middle of the range where the fraction was calculated.

As we see from the confusion matrix resulting from the analysis of the test set (Figure 20), a small fraction (12.3%, 557) of star-forming galaxies is classified as composite based on our diagnostic. Based on the previous discussion these objects have low overall sSFR resulting in weak star-forming activity signatures in the WISE bands. This discrepancy could be due to the fact that the optical-line based classification is based on SDSS spectra probing the nuclear regions of the galaxies, while our diagnostic is based on integrated photometry from the galaxies. In this case even weak nuclear star-forming activity may classify a galaxy as star-forming when SDSS spectra are considered, while in practice its near to mid-IR photometry is dominated by the older stellar population component.

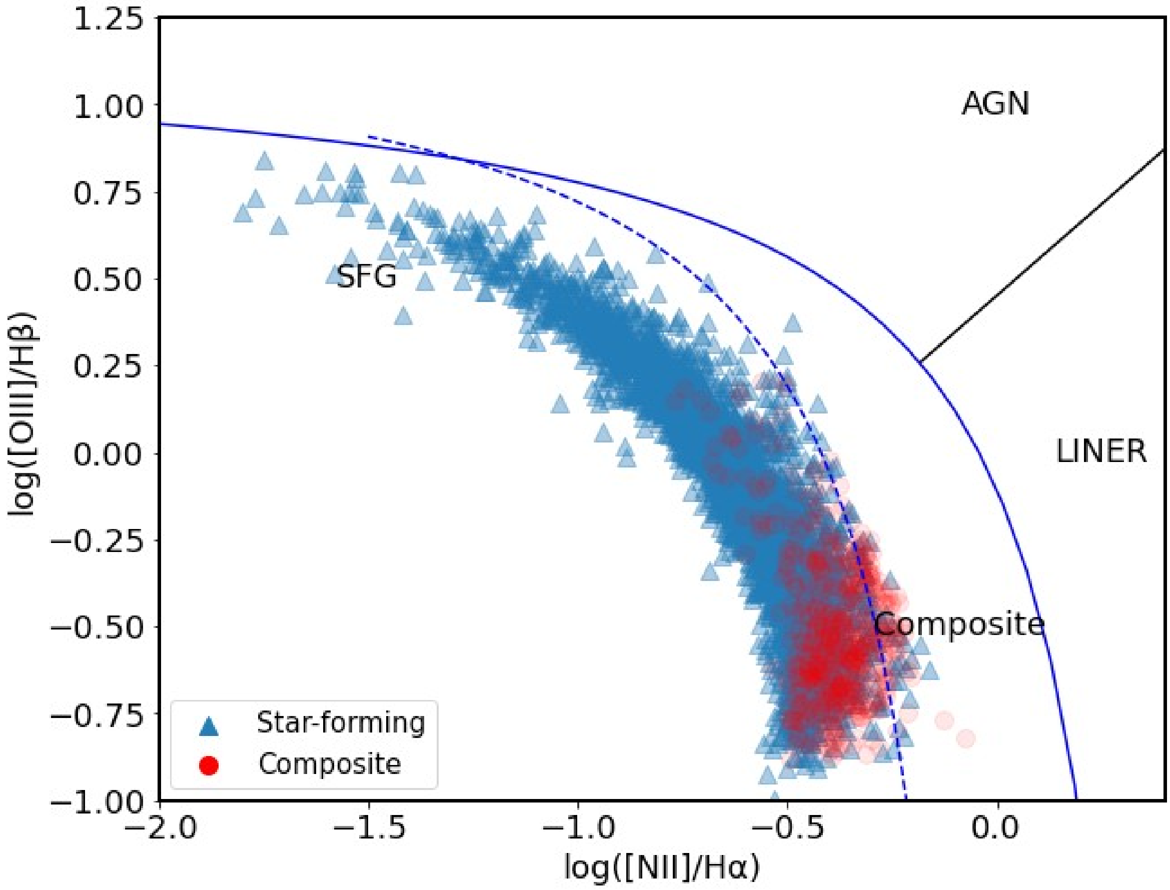


Figure 28: BPT (Baldwin et al. 1981) optical emission line plot of log([OIII]/Hb) against log([NII]/Hα). The 557 galaxies originally thought as star-forming but was classified by the new diagnostic as composites (red circles). The star-forming galaxies are presented with blue triangles. We can see that the bottoms of the SFG locus the galaxies tent to be classified by the new diagnostic as composite rather than SFG. The blue dashed line is the Kauffmann (Kauffmann et al. 2003) separating SFG and composites. The blue solid line is the Kewley line (Kewley et al. 2001).

Another possibility is that post-AGB stars in low sSFR galaxies have significant contribution in the photoionization of the interstellar medium. As shown in Byler et al. (2017, 2019), the increased contribution by hot evolved stars can produce optical line-ratios in the locus of composite or LINER (or LIER) objects. Indeed, as we see in Figure 29 the galaxies we classify as composites are found in the upper envelope of star-forming galaxies in the ([OIII]/Hb - [NII]/Ha) line ratio diagnostic. The fact that these galaxies do not fall entirely in the locus of composite galaxies could be due to the contribution by a weak star-forming component that contributes in the photoionization of the ISM.

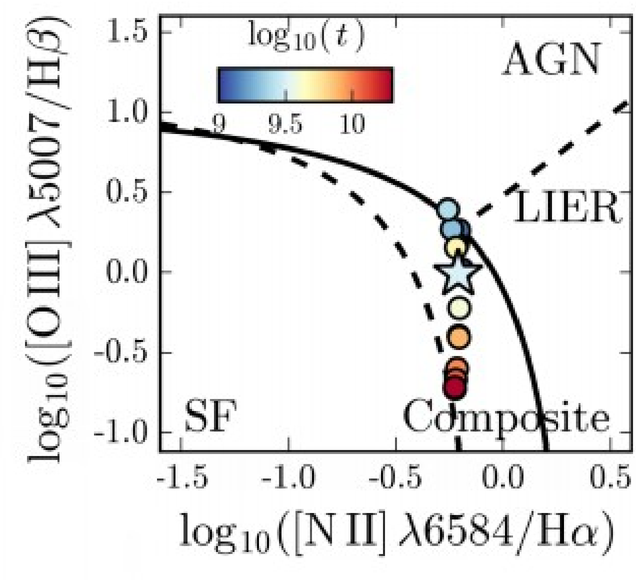


Figure 29: BPT plot of log([OIII]/Hβ) against log([NII]/Hα). The plot shows the post-AGB stars emission color-coded by age for ranges of 1 to 14 Gyr. Adopted from Fig. 8 of Byler et al. 2019.

4.5 Systematic behavior as function of g-r color

Another indicator of the dominant age of the stellar populations is their optical colors: bluer colors indicate younger stellar populations (e.g., Leitherer et al. 1999). The galaxies considered in our analysis are drawn from the SDSS spectroscopic sample and therefore, they have available high quality optical photometry in the SDSS bands. The SDSS survey also recorded photometry in 5 filters u, g, r, i and z (see section 2.3) for a large sample of galaxies. Here we consider photometry in the g-band (centered at 4686 Å) and in the r-band (centered at 6166 Å). Galaxies dominated by young hot stars will have negative and close to 0 g-r colors, while galaxies with older stellar populations will have higher values in their g-r colors. So, in order to understand better how the diagnostic actually works, in Figure 30, we plot the fraction of galaxies that were classified as star-forming by the multidimensional emission-line classification of Stampoulis et al. 2019 that the new diagnostic predicts to be in the different classes as function of g-r color.

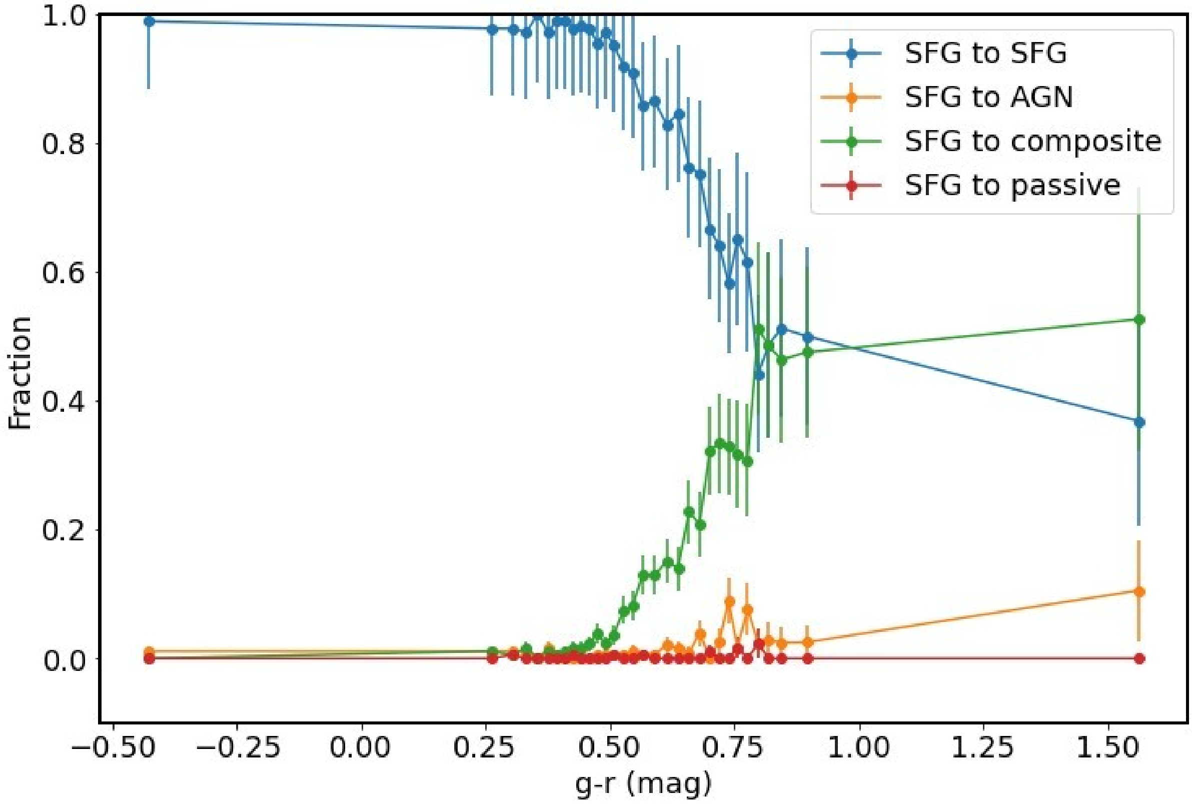


Figure 30: Plot of the fraction of galaxies as a function of g-r color. The fraction is calculated by dividing the number of galaxies that the diagnostic predicted to the in a particular class by the number of true star-forming (SFG) galaxies (as defined by Stampoulis et al. 2019) as function of g-r of SDSS in every bin. Each bin contains about 400 galaxies. Every point of the plot represents the middle of the range where the fraction was calculated.

In Figure 30 we see that the larger the value of the g-r color of a galaxy the higher the fraction of spectroscopically identified star-forming galaxies are predicted to be as composite by the Random Forest rather as star-forming. For smaller values of the g-r color the fraction of "true" star-forming that are predicted to be star-forming is 1. Both of these are as expected behaviors, as in star-forming galaxies we find that young populations of stars are dominant, while for a galaxy containing older populations has redder g-r colors. Also, higher g-r indicates lower sSFR as seen in the section 4.4.

In this section we tried to get an insight of the actual operation of the algorithm. We discovered that the 557 star-forming galaxies that were predicted as composite galaxies, actually had low sSFR indicating that this behavior is acceptable. We also saw that star-forming galaxies with redder g-r colors tent to be classified as composites rather than star-forming.

4.6 The case of LINERs

The confusion matrix showed that some AGN (which we note includes narrow and broad-line Seyfert galaxies, and LINERs) galaxies were classified as passive galaxies. Further investigation of these objects showed that most of them are actually LINERs. Their positions on the BPT diagrams, log([OIII]/Hβ) - log([NII]/Hα) and log([OIII]/Hβ) - log([OI]/Hα) in Figures 31 and 32 respectively.

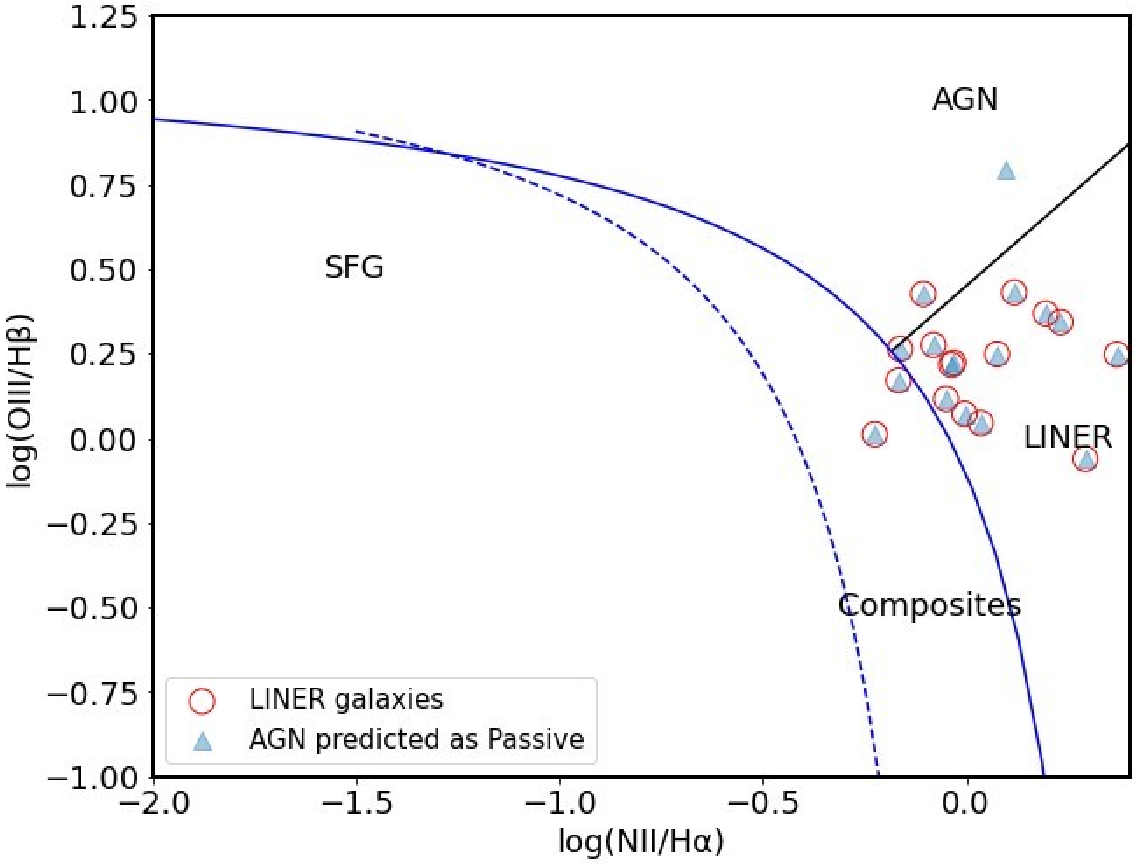


Figure 31: BPT (Baldwin et al. 1981) optical emission line plot of log([OIII]/Hb) against log([NII]/Hα) showing the AGN galaxies that predicted by the new diagnostic as passive galaxies (blue triangles). We mark with the red circles, the AGN galaxies that were predicted as passive but their optical line ratios classify them as LINERs. The blue dashed line is the Kauffmann line separating the star-forming from composites (Kauffmann et al. 2003). The blue solid line is a Kewley line separating composites from LINERs and AGN (Kewley et al. 2001). The black line between the Seyfert and LINERs is also presented in the plot (Schawinski et al. 07).

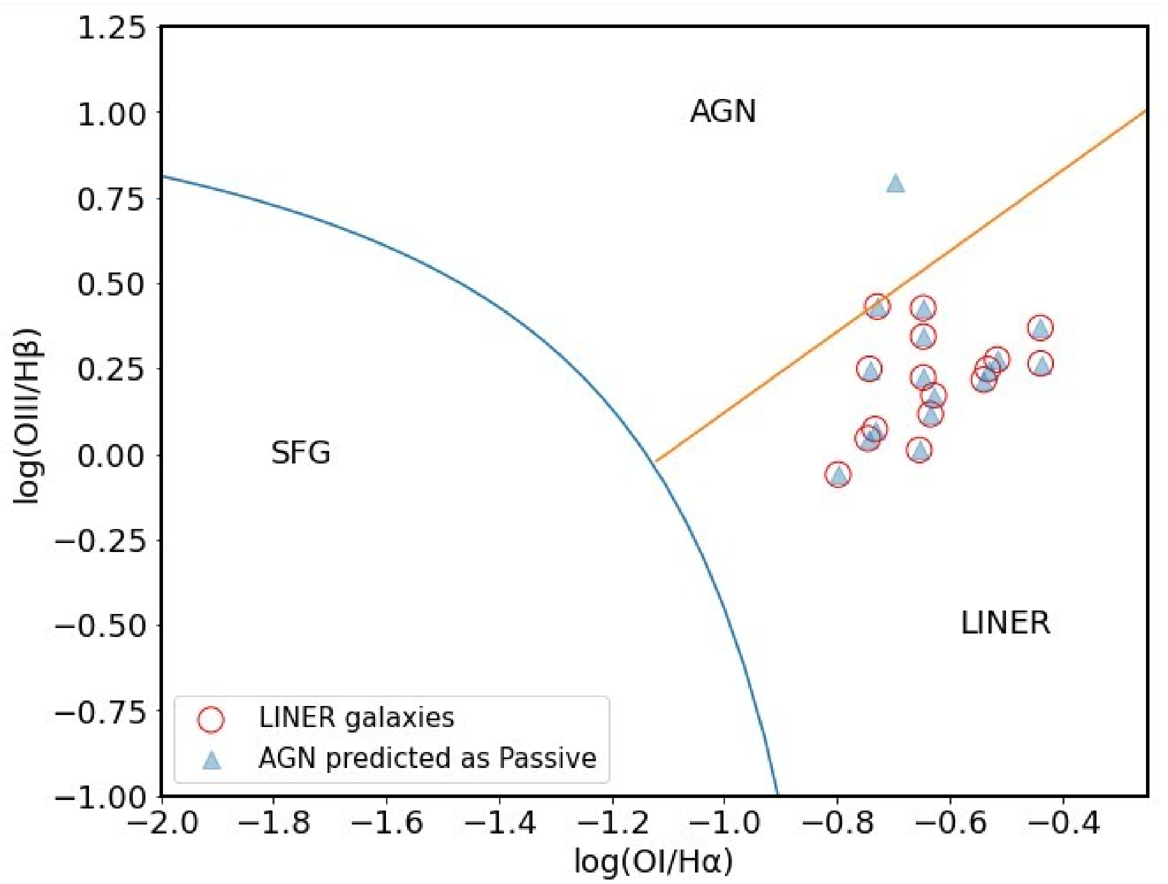


Figure 32: BPT (Baldwin et al. 1981) optical emission line plot of log([OIII]/Hβ) against log([OI]/Hα) showing the AGN galaxies that predicted by the new diagnostic as passive galaxies (blue triangles). For the AGN galaxies that were predicted as passive but their optical line ratios classify them as LINERs, we mark them with the red circles. The blue solid line is the Kewley et al. 2001 separating star-forming galaxies from the rest. Kauffmann line separating the star-forming from composites (Kauffmann et al. 2003). The orange solid line separates the AGN from the LINERs.

This is a very interesting result as the activity mechanism of some LINERs is considered to be emission from post-AGB stars and not from a central black hole (Singh et al. 2013). Passive galaxies have old stellar populations as well as post-AGB stars. This means that a passive galaxy can mimic an active one (Stasińska et al. 2008). Furthermore, it can do so, although LINERS are included in the AGN class during the training.

5 Classification of the full sample

5.1 Application of the diagnostic

After the training and testing our diagnostic tool, the 4-class diagnostic is used for the classification of the entire HECATE catalog. We apply our classification tool to galaxies with WISE photometry and with signal-to-noise above 3 in the three bands of interest (bands 1,2 and 3). The resulting sample is 138033 galaxies (68.7% of the overall catalog). The remaining sample either does not have available WISE photometry or the quality of their photometry is not adequate to provide reliable results (31.1% of the full sample). The percentages of the galaxy classes that are found in HECATE catalog after the application of the diagnostic are 45% star-forming, 23% AGN, 18% composite and 14 % passive galaxies.

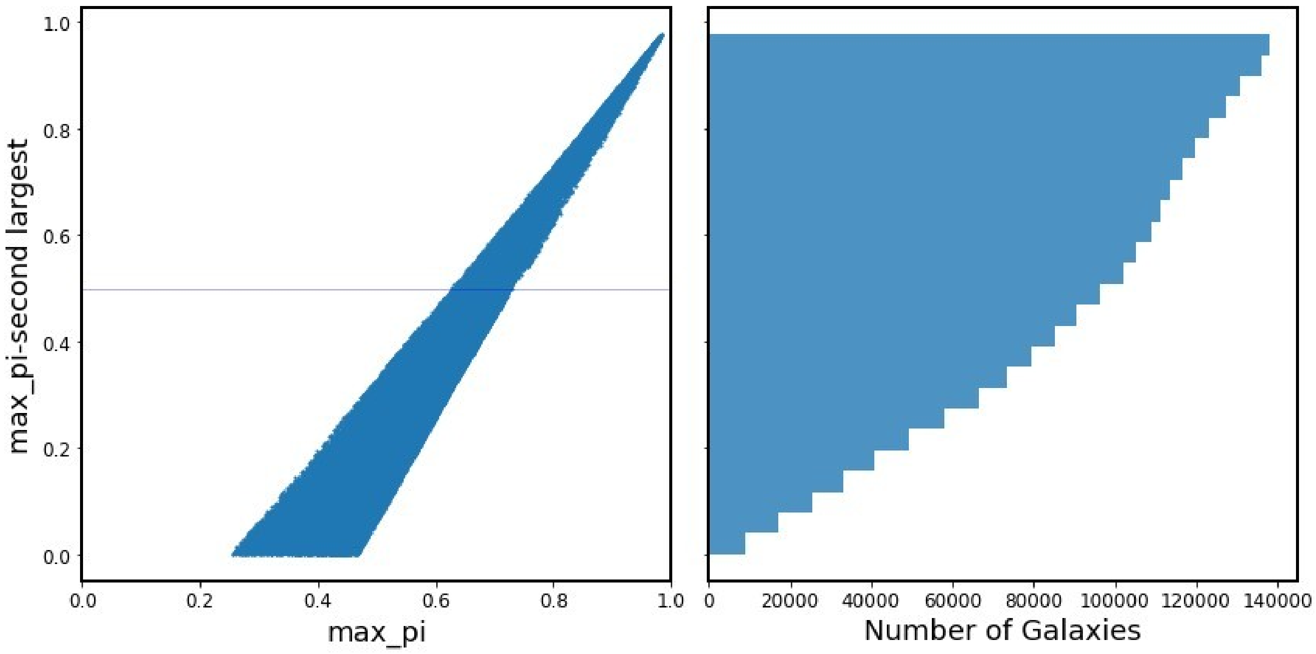


Figure 33: Left: A plot of the maximum minus the second largest probability against the maximum probability (max_pi) of the classification of each object. This plot represents the difference between the highest and the second-highest ranking class as predicted from the Random Forest for every object. On the right the cumulative plot of the same plot as on the left. We find that 14.6% of all galaxies have maximum probability (Random Forest predictions) above 75%.

A measure of reliability of the classification on the full HECATE catalog, is the number of objects that have been classified with maximum probability above a certain threshold (Figure 33). Two thresholds are chosen: 50% and 75%, the latter indicating high-confidence classifications. We find that 84% of the objects classified as star-forming have maximum probability above 50%, and 47% above 75%. For the class of the AGN 69% of the objects classified as AGN have maximum probability above 50% while no object seems to have maximum probability above 75%. These results, along with the statistics for the other classes are summarized in Table 7. We see that almost 50% of the star-forming galaxies are classified with high confidence. However, none of the AGN or Composite galaxies is classified with high confidence reflecting the mixing between the two classes seen in the confusion matrix analysis (section 4.1).

CLASS	Galaxies with maximum probability above 50% (%)	Galaxies with maximum probability above 75% (%)
Star-forming	84	47
AGN	69	0
Composite	53	0
Passive	68	5

Table 7: The reliability of the classification for each class after the application of the new diagnostic (Random Forest) in the full HECATE catalog.

It is instructive to compare the classification statistics for the test sample (for which we also have BPT diagnostics) and the full HECATE sample. The classification statistics for each class we consider for the two samples are shown in Table 8.

CLASS	Full HECATE composition (Random Forest) %	Test sample Random Forest composition %
Star-forming	45	65
AGN	23	8
Composite	18	12
Passive	14	15

Table 8: The composition of the HECATE catalog based on the four classes star-forming, AGN, composite and passive as classified by the Random Forest. On the right column, the percentages from the Random Forest in the test sample (all galaxies in the test sample have SDSS detections).

From Table 8 it is clear that there is some difference between the composition of the test sample (the one that has SDSS spectra) and the sample from full HECATE catalog. The passive galaxies have the same percentages in both samples. However, the fraction of star-forming galaxies drops in the full HECATE by ~20%, while the fraction of AGN increases by ~15%. This discrepancy is discussed further in section 5.3.

5.2 Distribution of HECATE galaxies

The three features used in the diagnostic (W2 luminosity, W1-W2 color and W2-W3 color), define a three-dimensional space as seen in Figure 34. A plot of all the galaxies in the HECATE catalog in this three-dimensional feature space shows the locus of each class and it can help us identify biases between the training and the full sample, interesting classes of objects, and further explore the behavior of the diagnostic. It is clearer that the full HECATE sample has the same distribution in each projections of this 3D space as the training data set. This ensures the applicability of the Random Forest classifier to the overall HECATE sample since the training sample covers the full range of each parameter and each class.

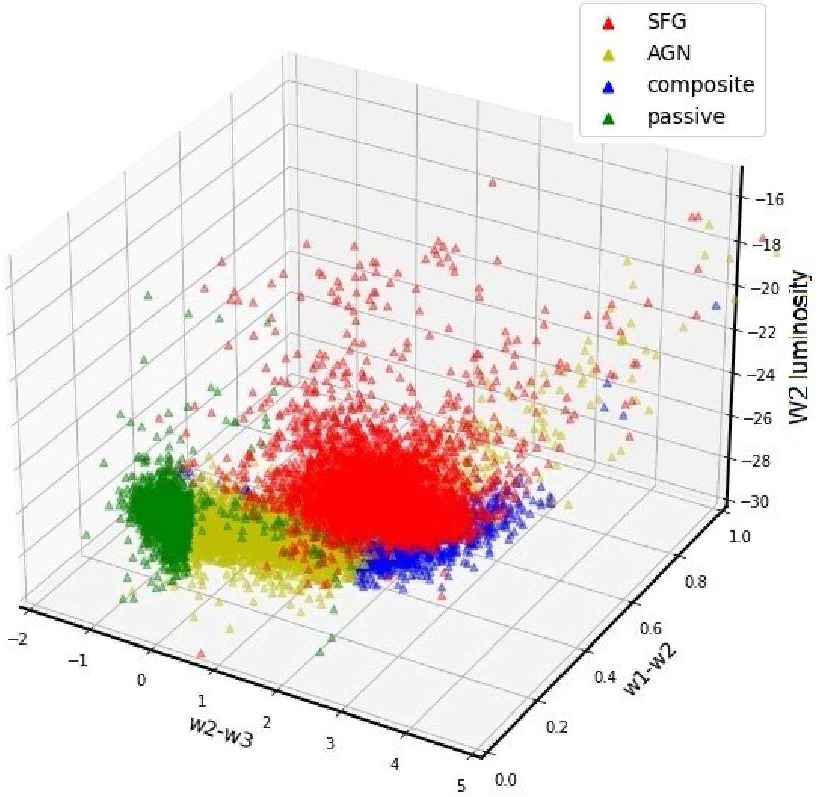


Figure 34: 3D diagram of the galaxies of HECATE catalog in the feature space. The classification of the objects is based on the new diagnostic.

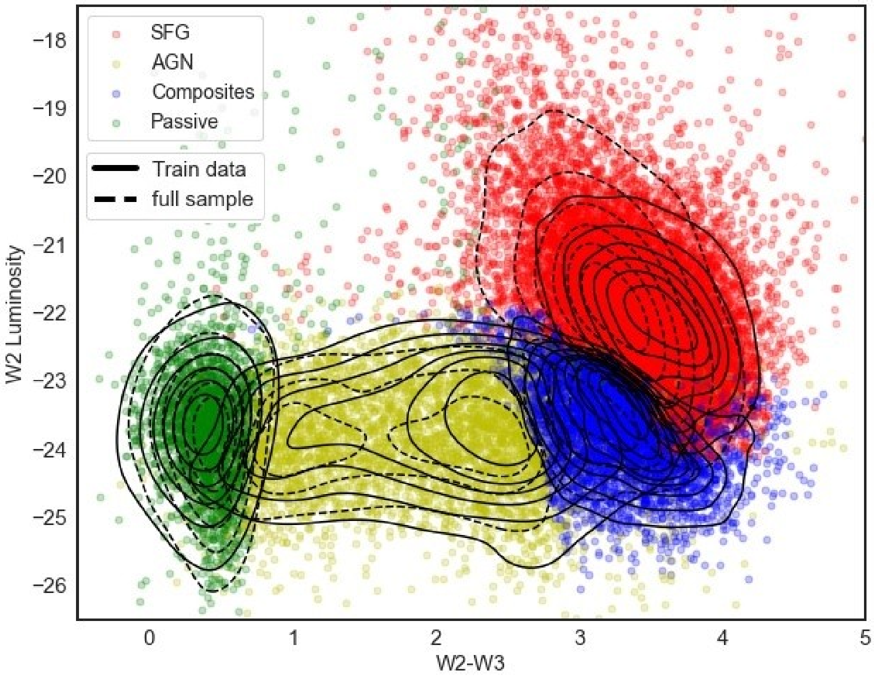


Figure 35: Plot of the W2 luminosity (absolute W2) against the W2-W3 color. This is a 2D projection of the 3D feature space. The labels of the galaxies have been assigned by the new diagnostic. The solid black contours correspond to the distribution of each class in the training set and the dashed black contours represents the full sample. It is clear that the full and the training set agree very well.

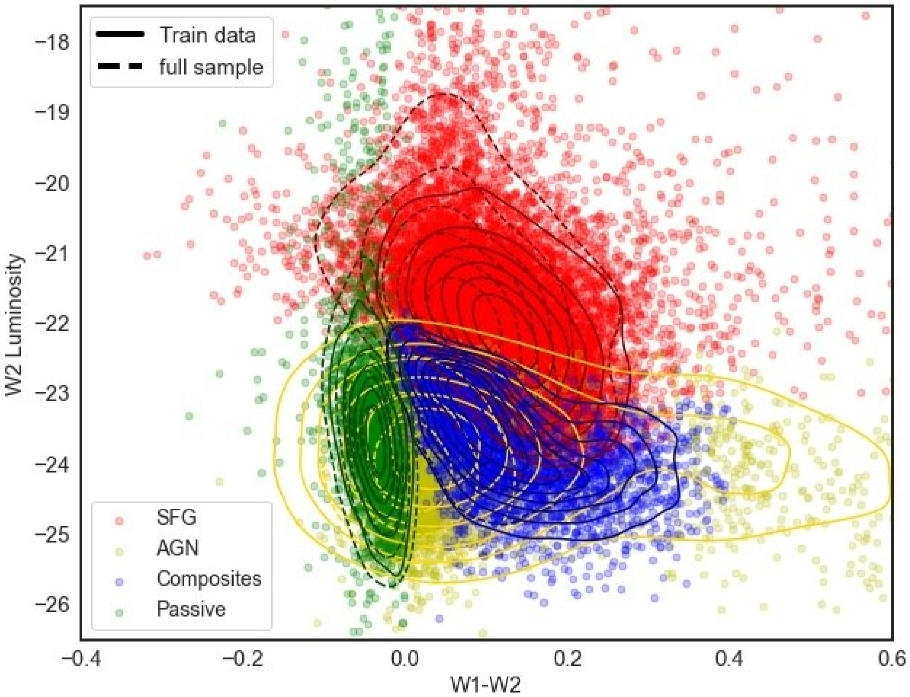


Figure 36: Plot of the W2 luminosity (absolute W2) against the W1-W2 color. This is a 2D projection of the 3D feature space. The labels of the galaxies have been assigned by the new diagnostic. The solid contours on every class represents the distribution of the training set and the dashed contours represent the full sample. It is clear that the full and the training set agree very well.

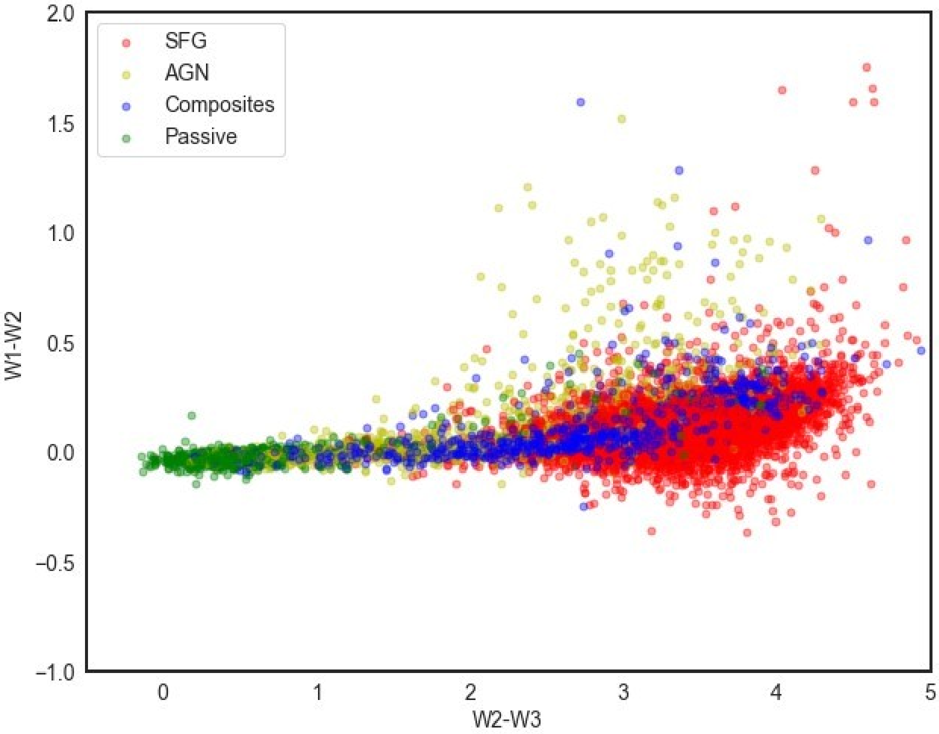


Figure 37: Plot of the W1-W2 against the W2-W3 (color-color diagram). A 2D projection of the 3D feature space. The labels have been assigned by the new diagnostic. The plot contains all the eligible galaxies for classification in the HECATE catalog.

Figures 35 and 36 show that the distribution of the full HECATE shows the same extent as the training set in each projection of the 3D space, despite that fact that it generally includes fainter objects and a much larger population of galaxies. Despite the fact that the training sample is drawn from the SDSS spectroscopic subset, it is representative of the overall population in this 3D space. Figure 37 represents the mixing that we find between the classes in the W1-W2 color against W2-W3 color projection.

5.3 AGN activity demographics

As mentioned in section 2.2, the HECATE catalog consists of all known galaxies in a volume of 200 Mpc radius, making it suitable for measuring of activity demographics in the local Universe. In this section we compare our activity demographics with those reported from other demographic studies in the local Universe. First, we consider the survey of a flux-limited (B_T < 12.5) sample of Northern galaxies performed by Ho et al. (1995, 1997e). This analysis showed that the 42% of all nearby galaxies are star-forming, 30% are AGN, 14% are composites and 14% passive. The application of our new diagnostic on the whole sample of

HECATE, gives 45% of star-forming galaxies, 23% AGN, 18% composite and 14% passive galaxies. These are very close to the results from the work of Ho et al. 1997e, despite the totally different selection of the two samples, the different classification methods, and the much larger (and lower luminosity) HECATE sample. This comparison, along with the demographics based on the SDSS spectroscopic sample are summarized in Table 9.

CLASS	Full HECATE composition (Random Forest) %	test sample (SDSS, Random Forest composition) %	Local Universe demographics (Ho et al. 1997e) %
Star-forming	45	65	42
AGN	23	8	30
Composite	18	12	14
Passive	14	15	14

Table 9: Comparison of percentage composition per class. First column: the results from the application of the Random Forest to the full HECATE catalog. Middle column: The predictions of the random Forest on the test sample which is a good measure of SDSS sample composition, as every galaxy in the test set has SDSS spectra. Right column: the demographics of the local Universe from the work of Ho et al. 1997e).

The contribution of each galaxy class in the training sample (SDSS) should represent the corresponding contribution for each class found in the local Universe. However, we find that, there is a discrepancy between the SDSS and the HECATE sample, with the SDSS containing a much larger fraction of star-forming galaxies (65% versus 45%) and a smaller fraction of AGN (8% versus 23%). This disagreement is due to the fact that the SDSS sample contains much fainter galaxies than the overall HECATE sample. Since star-forming galaxies tend to have lower luminosity than AGN galaxies, the SDSS sample is expected to be contain larger populations of star-forming galaxies. Although the HECATE sample contains all SDSS galaxies within a distance of 200 Mpc these comprise only 15% of the overall sample; the HECATE galaxies in the rest of the sky is based on shallower surveys which contain small fractions of the fainter star-forming galaxies.

6 Conclusions

The motivation behind this project was the definition of a new mid-IR diagnostic, as the existing diagnostics (Assef et al. 2013; Mateos et 2012) did not show the expected behavior when applied to the galaxies in the Heraklion Extragalactic Catalogue (Kovlakas et al. 2021). This discrepancy could be due to a number of reasons. Concerning the Assef et al. 2013 criterion, the reason may be that it only uses W1-W2 > 0.8 feature for AGN selection, when in our tool, we use more information by combining the W2 luminosity, and the (W1-W2) and (W2-W3) WISE colors. For the Mateos et al. 2012, the explanation for our differences maybe that W2 luminosity helps a lot in the discrimination of the four class considered here.

Concerning the performance of our new diagnostic, we find that it shows excellent performance in discriminating star-forming and passive galaxies in terms of overall scores and high classification confidence. On the other hand, it is less efficient in discriminating between AGN and composite galaxies. This is reflected on the poor performance for the composite galaxies and the relatively low confidence (highest prediction probability, see section 4.2) of the diagnostic when classifying a galaxy as AGN or composite. The most likely explanation for this limitation is the extensive mixing between these two classes (Figure 37). We will try to address this limitation in the future by adding more features that are key (optical, UV, X-rays) for the discrimination between AGN and composite galaxies.

In order to investigate the effect of the strong bias of the original sample toward star-forming galaxies we followed two methods that mitigate sample imbalance: (a) random removal of the excess objects, and (b) up-sampling (by means of simulation from multi-dimensional Gaussian) of the under-samples types of objects. Neither of these two methods gave any different results from each other or the analysis with the original sample. The magic number (section 2.5) does not fix imbalance but it allows the use of photometry for very faint objects (especially for passive galaxies).

Conclusions and results derived from this project are:

- This new diagnostic tool is able discriminate between the four classes of galaxies star-forming, AGN, composite and passive galaxies by their mid-IR spectrum with reasonable precision. It performs well on star-forming and passive galaxies (precision ~95%) and moderately on AGN.

- The mid-IR colors are sensitive to the star-formation process as shown by the Random Forest classification results. Furthermore, our analysis is sensitive to star-forming galaxies with even relatively low specific SFR ($\log_{10}$(sSFR)=-10 1/yr).

- A small fraction (12.3%) of star-forming galaxies are classified as composites. The fact that these are predominantly galaxies with redder optical colors and very low specific SFR, suggests that they may have a significant populations of post-AGB stars contributing in their ionization.

- The percentage of the galaxies in the HECATE catalog that now have classification increased from 31.7% to 68.7%.

- The demographics of the HECATE catalog are very similar to the demographics that are predicted in the local Universe from other studies (e.g., Ho et al. 1997e)

Acknowledgements

I want to thank professor Andreas Zezas for giving me the chance to work with him on this Astrophysics project, for all the knowledge that I gained in the process and for his suggestions on the structure of this thesis. Also, I want to thank Elias Kyritsis for our discussions which expanded my knowledge on the Machine Learning methods and for his valuable suggestions on the writing of this thesis and Kostas Kouroubatzakis for his valuable contribution on the interpretation of the results. This publication makes use of data products from the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration. This publication makes use of data from the Sloan Digital Sky Survey III. Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University,

the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

References

Alonso-Herrero, A., et al. 2014, MNRAS 443, 2766, doi: 10.1093/mnras/stu1293

Assef, R. J., Stern D.,Kochanek C.S.,et al. 2013, The Astrophysical Journal, 772, 26, doi: doi:10.1088/0004-637X/772/1/26

Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, PASP, 93, 5, doi: 10.1086/130766

Brinchmann J., et al. 2004, MNRAS, 351, 1151

Brown, M.J.I.,et al. 2014, The Astrophysical Journal Supplement Series, 212, 18, doi: 10.1088/0067-0049/212/2/18

Byler, N., Dalcanton, J.J., Conroy, C., et al. 2019, The Astronomical Journal, 158, 2, doi: 10.3847/1538-3881/ab1b70

Byler, N., Dalcanton, J. J., Conroy, C., & Johnson, B. D. 2017, ApJ, 840, 44, doi: 10.3847/1538-4357/aa6c66

Cortes C., Vapnik V., 1995, Machine learning, 20, 273

Donley J.L., et al., 2012, ApJ 748, 142, doi: 10.1088/0004-637X/748/2/142

Genzel R., 2014, Nature 391, 17, doi: 10.1038/34029

Lang D., Hogg D. W., Schlegel D. J., 2016, AJ, 151, 36

Ho, L.C., Filippenko, A.V., Sargent, W.L.W., Peng, C.Y. 1997e. Ap. J. Suppl. 112:391

Ho, L. 2008, Annual Reviews of Astronomy and Astrophysics,46, 476, doi: 10.1146/annurev.astro.45.051806.110546

Mateos S., Alonso-Herrero A., Carrera F.J,et al. 2012, MNRAS, 426, 3271, doi:10.1111/j.1365-2966.2012.21843.x

Kauffmann, G., Heckman, T. M., Tremonti, C., et al. 2003, MNRAS, 346, 1055, doi: 10.1111/j.1365-2966.2003.07154.x

Kauffmann G., et al., 2003, MNRAS, 346, 1055

Kewley, L. J., Dopita, M. A., Sutherland, R. S., Heisler, C. A., & Trevena, J. 2001, ApJ, 556, 121, doi: 10.1086/321545

Koekemoer A. M., et al., 2011, ApJS, 197, 36

Kovlakas, K., Zezas, A., Andrews, J.J.,et al. 2021, MNRAS 506, 1896 doi: 10.1093/mnras/stab1799

Leitherer, C., Schaerer, D., et al. 1999, The Astrophysical Journal Supplement Series, 123, 3,doi: 10.1086/313233

Salim, S., Lee, J.C., Janowiecki, S., et al. 2016, The Astrophysical Journal Supplement Series, 227, 2, doi: 10.3847/0067-0049/227/1/2

Schawinski, K., Kaviraj, S., Khochfar, S., et al. 2007, ApJS, 173, 512

Singh, R., van de Ven G., Jahnke K., et al. 2013, A&A 558, A43

Skrutskie M. F., et al., 2006, AJ, 131, 1163

Stampoulis, V., van Dyk D.A, Kashyap, V.L., Zezas, A., MNRAS 485, 1085, doi:10.1093/mnras/stz330

Stasinska, G., Asari, N. V., et al. 2008, MNRAS 391, 29 doi: 10.1111/j.1745-3933.2008.00550.x

Stern D., et al., 2005, ApJ, 631, 163 doi: 10.1086/432523

Tremonti C. A., et al., 2004, ApJ, 613, 898

Werner, M. W. et al. 2004, ApJS, 154, 1

Wright, E.L., Eisenhardt, P.R, Mainzer, A.K, et al. 2010, The Astronomical Journal, 140, 1868, doi:10.1088/0004-6256/140/6/1868

York, D. G., et al., 2000, AJ, 120, 1579, doi: 10.1086/301513

Appendix A – Validation curves

Below are the validation curves used in combination with the grid search algorithm for the optimization of the Random Forest.

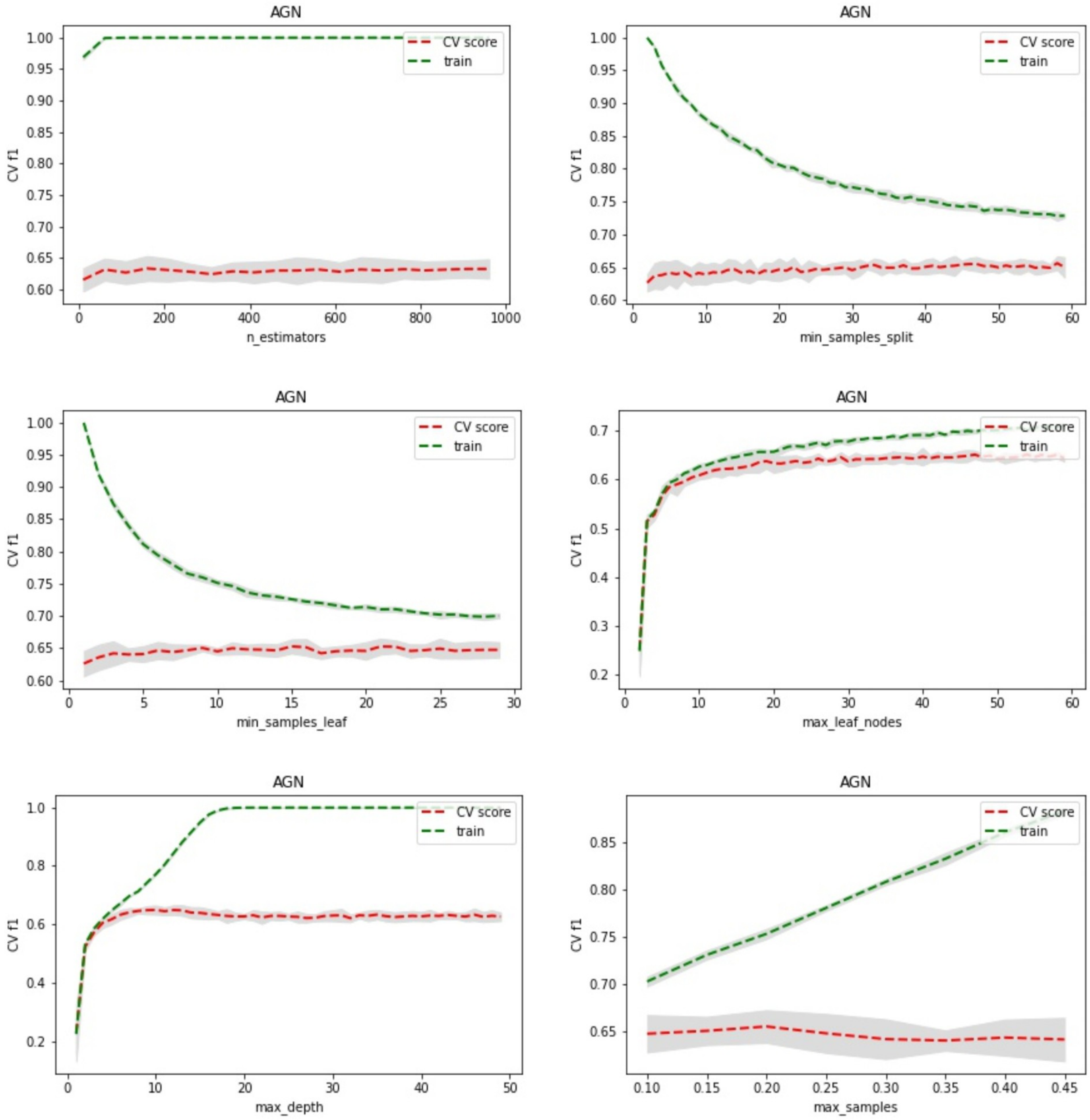


Figure 38: Validation curves of the Random Forest performance on the AGN class. In the plots above, f1-score as a function of each optimizable hyperparameter is presented. Each plot represents the performance for a different hyperparameter. The red dashed line represents the score on the training data set and the green one with the 5-fold cross-validation method. The shaded area represents the uncertainty (standard deviation) on the 5-fold cross-validation.

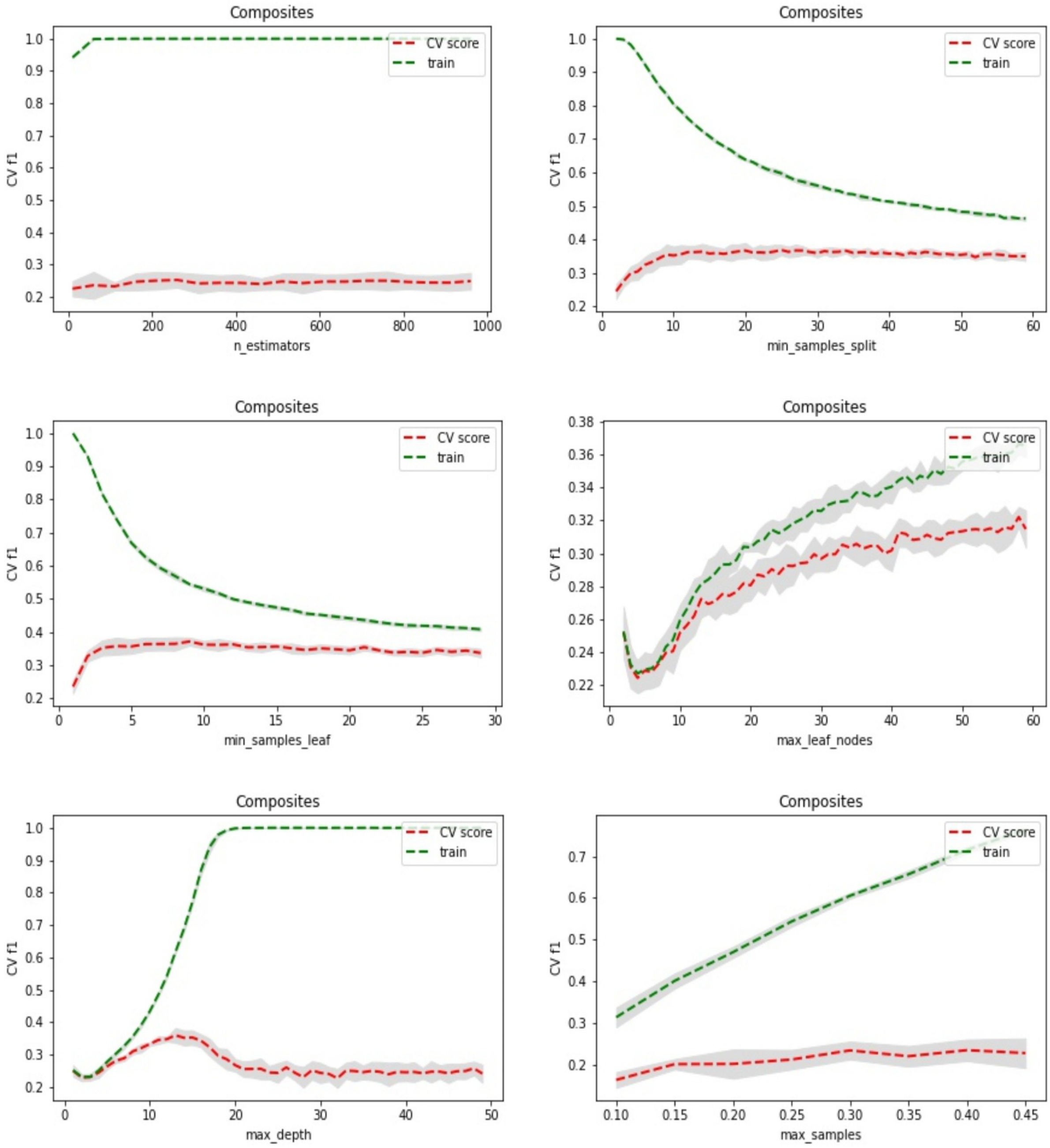


Figure 39: Validation curves of the Random Forest performance on the composite galaxies. In the plots above, f1-score as a function of each optimizable hyperparameter is presented. Each plot represents the performance for a different hyperparameter. The red dashed line represents the score on the training data set and the green one with the 5-fold cross-validation method. The shaded area represents the uncertainty (standard deviation) on the 5-fold cross-validation.

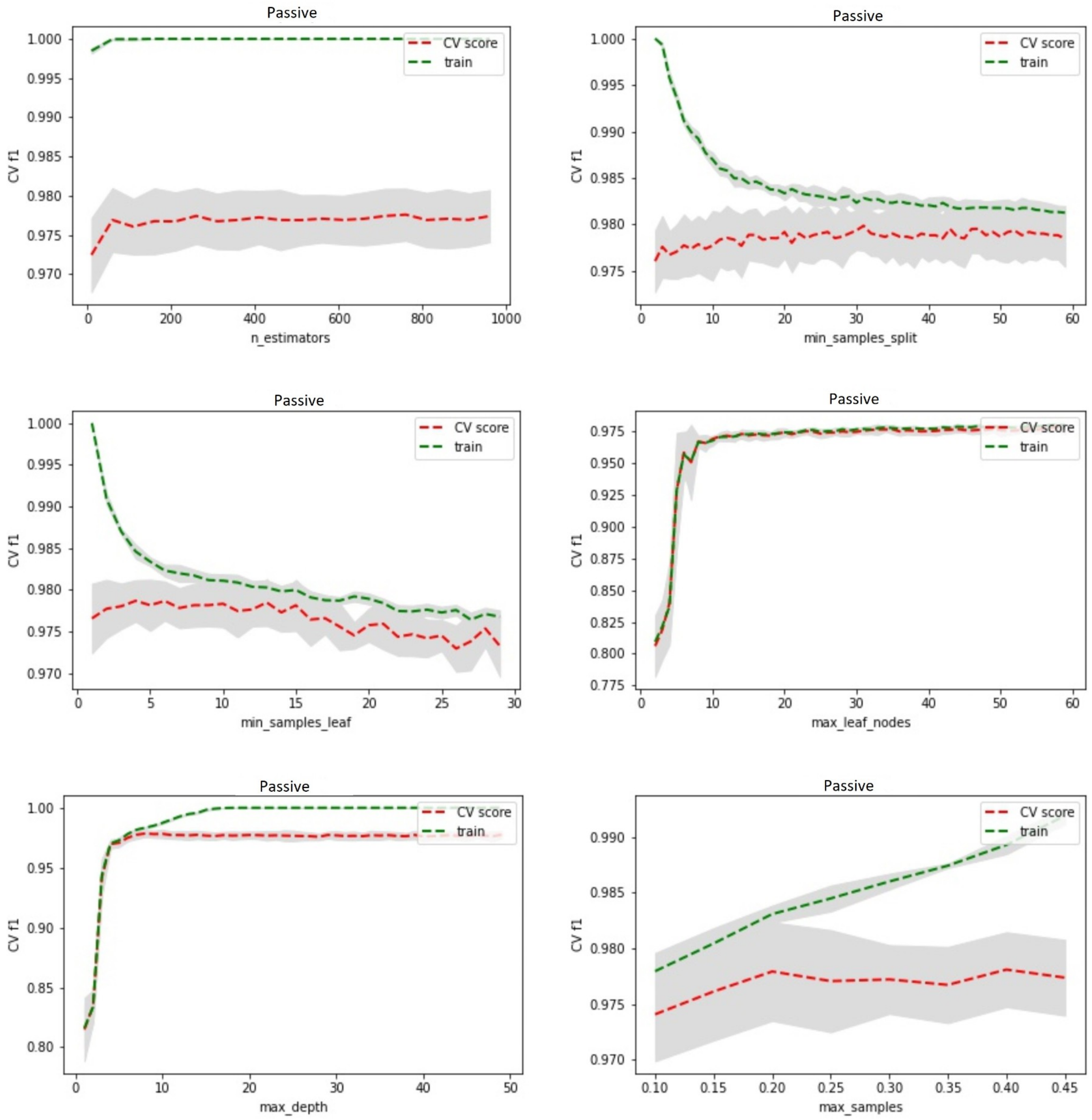


Figure 40: Validation curves of the Random Forest performance on the passive galaxies. In the plots above, f1-score as a function of each optimizable hyperparameter is presented. Each plot represents the performance for a different hyperparameter. The red dashed line represents the score on the training data set and the green one with the 5-fold cross-validation method. The shaded area represents the uncertainty (standard deviation) on the 5-fold cross-validation.

Appendix B – The 3-class diagnostic

B.1 The need of the 3-class diagnostic

In section 3, we described the definition of a 4–class diagnostic based on the WISE hybrid photometry scheme considering star-forming, AGN, composite, and passive galaxies. Although the performance of this diagnostic for the star-forming and passive galaxies was excellent, there was significant mixing between the composite and AGN galaxies, and to a less extent between composite and star-forming galaxies. The above observation led to the conclusion that a 3-class diagnostic might be more convenient. This 3-class diagnostic separates galaxies into 3-classes: star-forming, passive and "other" (or other emission-line object) galaxies. The class of "other", contains the galaxy classes of LINERs, AGN and composites. The data and photometry scheme used for the definition of this diagnostic, are the same used for the 4-class diagnostic (section 2). The distribution of galaxies for the 3-class diagnostic in the 3D feature space is presented in the Figure 41. The purpose of this 3-class diagnostic is to discriminate with high precision between star-forming and passive galaxies. Further discrimination between AGN and composite may be possible by adding more features.

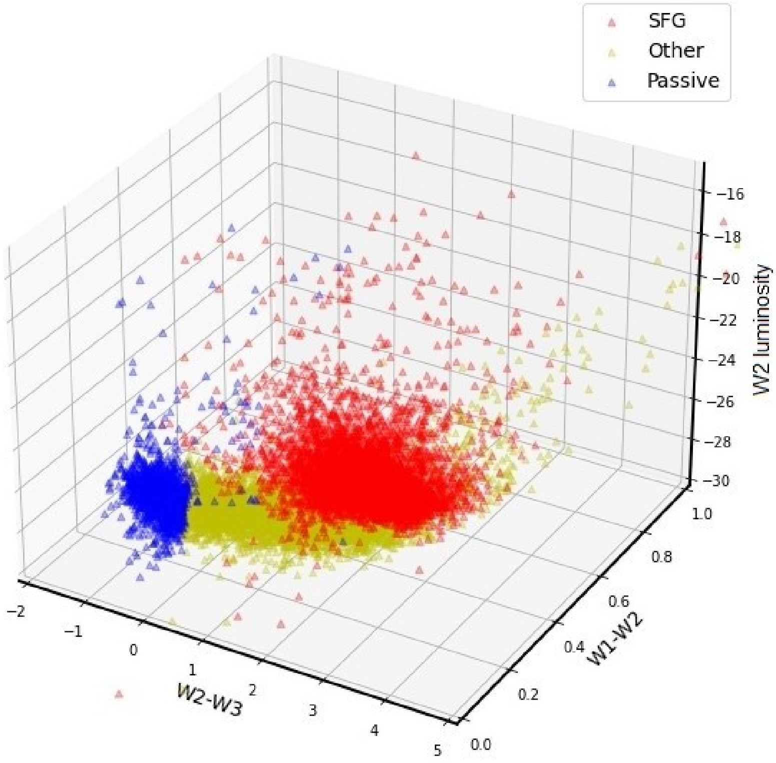


Figure 41: The distribution of the three classes in the 3D feature space. Star-forming in red, Other (LINER + Seyfert + composite) in yellow and passive in blue.

The algorithm used for the definition of this 3-class diagnostic is the Random Forest (section 3.1). The procedure for the training and the optimization of the algorithm was similar to the one used for the 4-class diagnostic tool. The data selected with the same criteria (section 2.5) and were split into training and test set with 70%-30% proportion respectively. The optimization procedure followed for the 3-class diagnostic included the same hyperparameters (section 3.4): max_depth, max_leaf_nodes, max_samples, min_samples_leaf, min_samples_split and n_estimators. Furthermore, the value for the bootstrap hyperparameter is set to 'True', the class_weight to 'balanced_subsample' due to high imbalance of the data, and the criterion hyperparameter to 'gini'. The best values are in the Table 10.

Hyperparameter	**Best Value**
n_estimators	250
max_leaf_nodes	34
max_depth	15
max_samples	0.1
min_samples_leaf	16
min_samples_split	39
bootstrap	True
class_weight	balanced_subsample
criterion	gini

Table 10: The best hyperparameters for the 3-class diagnostic for the Random Forest based on the grid search algorithm.

B.2 Performance metrics

Following the same process as in section 3, the confusion matrix (Figure 42), recall, precision and f1-score (Table 11) are calculated as a measure of performance. These metrics, were calculated on the test sample. In order for a diagnostic to be reliable, it has to make predictions that have high confidence. In this case, a measure of confidence, is the difference between the maximum probability and the second largest probability of a prediction. As done before in the section 4.2, for the 4-class diagnostic. The plot of difference between the maximum and the second largest probability against the maximum probability is presented in Figure 43. As in section 4.1 the confusion matrix can give us valuable information about the performance of the diagnostic (Figure 42). Also, the performance metrics (precision, recall and f1-score) are presented in Table 11.

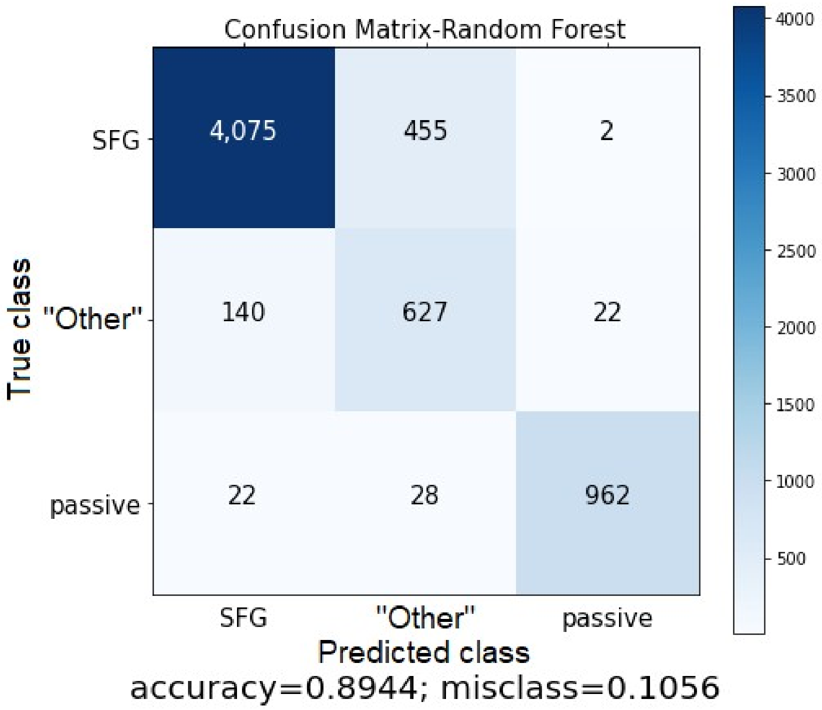


Figure 42: Confusion matrix for the 3-class Random Forest diagnostic.

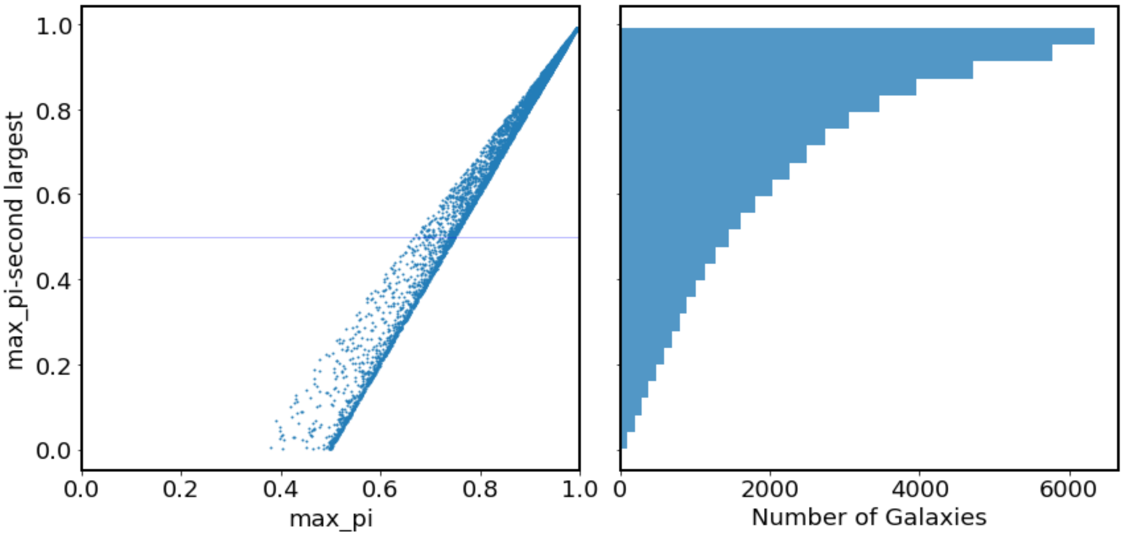


Figure 43: On the left the plot of the maximum minus second largest probability against the maximum probability (max_pi) predicted for all objects to belong to a particular class. On the right the cumulative plot of the same plot introduced on the left. We find that 98.8% of the objects present

maximum probability above 50%, and 75.6% above 75%. The galaxies used for these plots are from the test set, including all three classes.

	Precision	Recall	F1-score
Star-forming	0.96	0.90	0.93
“other”	0.56	0.79	0.66
Passive	0.98	0.95	0.96

Table 11: Performance matrix, precision, recall and f1-score for the 3-class Random Forest diagnostic. The class “Other” translates as other emission-line objects and it contains objects from AGN, composite and LINER galaxies according the classification of Stampoulis et al. 2019.

B.3 Discussion on the 3-class diagnostic results

We see that comparing these probability plots (Figure 43) with the ones for the four-class diagnostic (section 4.2), these predictions have more certainty as the Random Forest assigns higher probabilities for its predictions. Comparing the performance metrics of the 3-class diagnostic (Table 11) with the performance metrics of the 4-class diagnostic (section 4.1; Table 5) we can see that the precision and recall of the star-forming and passive galaxies show a slight improvement.

There are 140 galaxies predicted as star-forming while their true class was "Other" (i.e. AGN, composite, or LINER). From them, 98 are composites and 42 are AGN. By further investigation, these 42 AGN present low Hα luminosity (~10^{39} erg/sec, Figure 44). The fact that these AGN have low Ha luminosity indicates that they are low luminosity AGN and their emission can be diluted by the emission of the host galaxy in the WISE bands.

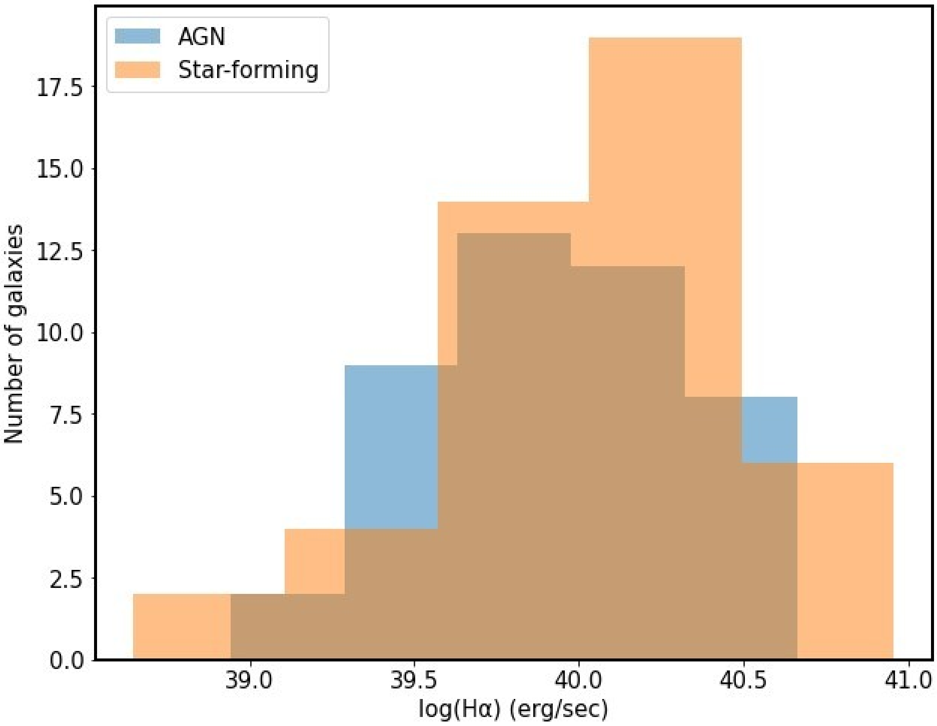


Figure 44: Histogram of Hα luminosity for AGN predicted to be star forming galaxies. We can see that the Hα luminosity of the AGN have similar distribution as the star-forming galaxies. The result must be taken cautiously as the sample is small (~50 galaxies).

Also, there are some star-forming galaxies, about 10% or 455 objects, that this 3-class diagnostic predicted as "other". To find the explanation why this is happening we will plot the fraction of star-forming galaxies that are classified as "others" as function of sSFR

(Figure 45). The sSFR values were derived according to the work of Salim et al. 2016 (section 4.4).

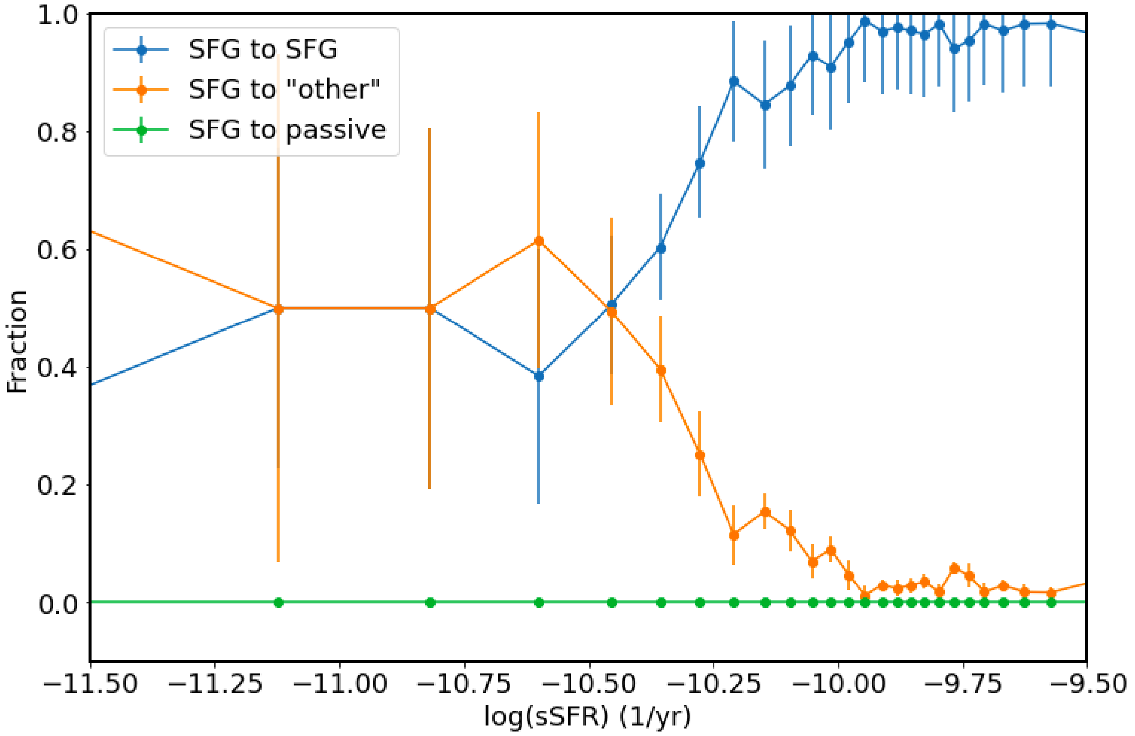
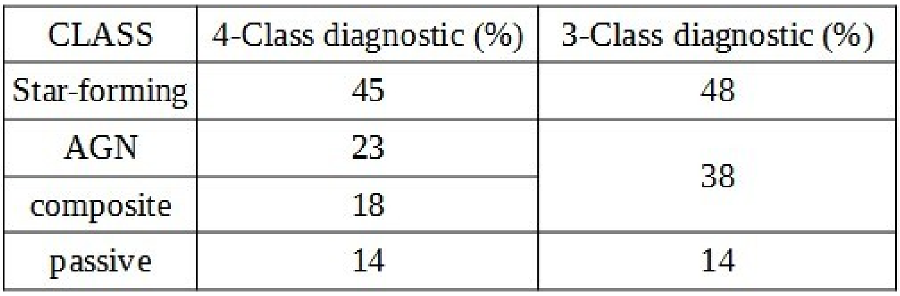


Figure 45: Plot of fraction of galaxies as a function of specific star formation rate log(sSFR). The fraction is calculated by dividing the number of galaxies that the diagnostic predicted to the in a particular class by the number of “true” star-forming galaxies (as defined by Stampoulis et al. 2019) as function of log(sSFR).

Figure 45 shows the same behavior as the 4-class diagnostic in section 4.4. As the sSFR of a galaxy drops, the more likely it is to be predicted as “other”.

Applying the 3-class diagnostic in the full HECATE catalog, the composition per class is found to be: 48% star-forming, 38% “other” and 14% passive. These results are presented in the Table 12 and compared with the results found after the application of the Random Forest 4-class diagnostic of the full HECATE catalog (section 5.1;Table 8).



CLASS	4-Class diagnostic (%)	3-Class diagnostic (%)
Star-forming	45	48
AGN	23	38
composite	18	
passive	14	14

Table 12: Comparison of the percentages of the classes found in the 3-class and 4-class diagnostic. In the third column contains the result of the classification by the 3-class diagnostic. The class "other" is the combination of AGN and composite galaxies.

The general result after comparing the two diagnostics is that they behave generally the same but is a bit more reliable (see section B.2) when predicting star-forming and passive.

Appendix C – Decision boundaries for the 4-class diagnostic

Due to the nature of the Random Forest algorithm, the only way that one can have the result of the classification for a galaxy is to use the algorithm in a Python computer program environment. For convenience purposes, we fit a Support Vector Machine or SVM (Cortes & Vapnik 1995) in order to have the mathematical equations of the decision boundaries.

The SVM is a supervised machine learning algorithm. The discrimination in the different classes is achieved by support vectors between the data points of every class. The target of the algorithm is to minimize the distance of that support vectors with the lowest possible mixing between the different classes. That particular property of the algorithm allows us to extract the mathematical equations of the boundaries.

In section 4, we analyzed the results of the Random Forest classification. It was shown that even though some galaxies changed classification after the application of the new diagnostic, that change was justified by their properties (sSFR and g-r color). With that observation in mind, the result of the classification performed by the new diagnostic is set as true target labels for classification on the SVM algorithm. In that way, we do not seek to compare the performance of the two algorithms (Random Forest against SVM), instead we try to define the boundaries of the four classes in the 3-Dimensional feature space where the labels were defined by Random Forest. In other words, the target labels for the SVM were provided by the new diagnostic. The SVM algorithm was provided by the scikit-learn Python library. The kernel that was chosen was rbf (radial Gaussian kernel) as it was the only one from the kernels available that was able to describe the complexity of our problem.

The equations of the boundaries between the four classes are the following:

Star-forming:

- W2_luminosity - $0.13(W2\text{-}W3)^2$ - $0.73(W1\text{-}W2)^2$ - 1.83(W2-W3)(W1-W2) + 2.65(W2-W3) + 6.96(W1-W2) > - 15.47

AGN:

- W2_luminosity + $1.31(W2\text{-}W3)^2$ - $8.25(W1\text{-}W2)^2$ - 1.62(W2-W3)(W1-W2) - 4.58(W2-W3) + 4.68(W1-W2) < -25.85

- W2_luminosity + $1.12(W2\text{-}W3)^2$ - $43.25(W1\text{-}W2)^2$ + 9.23(W2-W3)(W1-W2) - 5.11(W2-W3) - 32.81(W1-W2) < -27.03

Composites:

- W2_luminosity - $0.13(W2\text{-}W3)^2$ - $0.73(W1\text{-}W2)^2$ - 1.83(W2-W3)(W1-W2) + 2.65(W2-W3) + 6.96(W1-W2) < - 15.47

- W2_luminosity + $1.31(W2\text{-}W3)^2$ - $8.25(W1\text{-}W2)^2$ - 1.62(W2-W3)(W1-W2) - 4.58(W2-W3) + 4.68(W1-W2) > -25.85

- (W2-W3) > 2

Passive:

- W2_luminosity + $1.12(W2\text{-}W3)^2$ - $43.25(W1\text{-}W2)^2$ + 9.23(W2-W3)(W1-W2) - 5.11(W2-W3) - 32.81(W1-W2) > -27.03

- (W2-W3) < 2.