ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

Galaxy activity classification and dominant photo-ionization mechanism characterization using optical spectra and machine learning methods

by

Charalampos Daoutis

Submitted in Partial Fulfillment of the

Requirements for the Master's Degree

in Advanced physics

Supervised by Prof. Andreas Zezas

School of Sciences and Engineering
Department of Physics

University of Crete

Heraklion, Greece

September, 2022

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Prof. Andreas Zezas, for giving me the opportunity to work with him in a such interesting and important topic and for the knowledge I acquired from him during this journey. His guidance and continuous support were extremely valuable for the fulfilment of this thesis. Furthemore, I also want to acknowledge Elias Kyritsis, Dr. Paolo Bonfini and Dr. Kostas Kouroumpatzakis, who besides their busy schedule, they always spared some time to share their knowledge and advise me. Also I want to thank Dr. S. Salim, for providing as with valuable data that made this project possible. Last but not least, I want to thank my family who are always there to support me and for helping me to reach this far.

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions.

SDSS-IV acknowledges support and resources from the Center for High Performance Computing at the University of Utah. The SDSS website is www.sdss.org.

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

Περίληψη

Οι μέθοδοι ταξινόμησης για τον χαρακτηρισμό της δραστηριότητας ενός γαλαξία έχουν μεγάλη σημασία στην παρατηρητική αστροφυσική. Παρόλο που πολλά διαγνωστικά εργαλεία έχουν κατασκευαστεί τα τελευταία χρόνια, η συντριπτική πλειονότητά τους αφορά μόνο γαλαξίες που παρουσιάζουν γραμμές εκπομπής ή είναι εξειδικευμένα μόνο σε μία κατηγορία δραστηριότητας (π.χ., ενεργοί γαλαξίες). Επιπλέον, σχεδόν κανένα από αυτά δεν είναι σε θέση να συμπεριλάβει όλους τους πιθανούς τύπους δραστηριότητας σε ένα ενιαίο σχήμα. Επιπλέον, αποτυγχάνουν να αντιμετωπίσουν σωστά το ζήτημα των γαλαξιών που έχουν σύνθετη δραστηριότητα. Σε αυτή την εργασία, σκοπεύουμε να ορίσουμε ένα διαγνωστικό εργαλείο βασισμένο σε μεθόδους μηχανικής μάθησης λαμβάνοντας υπόψη τρεις κατηγορίες που είναι αντιπροσωπευτικές των κύριων μηχανισμών ιονισμού των αερίων: σχηματισμός νέων άστρων, ενεργοί πυρήνες και διέγερση από παλαιούς αστρικούς πληθυσμούς που υπάρχουν κυρίως σε ελλειπτικούς γαλαξίες. Για το σκοπό αυτό, εκπαιδεύουμε έναν αλγόριθμο Τυχαίου Δάσους που χρησιμοποιεί συνολικά τέσσερα χαρακτηριστικά. Τρία από αυτά είναι τα ισοδύναμα πλάτη των φασματικών γραμμών του υδρογόνου, των απαγορευμένων γραμμών του αζώτου και οξυγόνου, που βρέθηκε ότι παρέχουν εξαιρετική διακριτική ισχύ για τους τρεις κύριους τύπους δραστηριότητας που μπορούν να βρεθούν σε έναν γαλαξία. Το τέταρτο χαρακτηριστικό είναι ένας δείκτης της μέσης ηλικίας των αστρικών πληθυσμών. Καταφέρνουμε να επιτύχουμε ακρίβεια $\sim$99%. Λόγω της υψηλής απόδοσης που επιτεύχθηκε στις κυρίες κατηγορίες δραστηριότητας και με βάση τις προβλεπόμενες πιθανότητες που παρέχονται από το Τυχαίο Δάσος, μπορούμε να εφαρμόσουμε αυτή τη μέθοδο στις κατηγορίες γαλαξιών σύνθετης δραστηριότητας προκειμένου να προσδιορίσουμε την κυρίαρχη πηγή διέγερσης του αερίου σε αυτούς. Για αυτόν τον λόγο, αυξάνουμε επίσης τις διαθέσιμες κατηγορίες δραστηριότητας για να παρέχουμε εκλεπτυσμένες προβλέψεις για τις κλάσεις συνθέτης δραστηριότητας. Επομένως, εκτός από τις κυρίες τάξεις δραστηριότητας, προσθέτουμε τις κατηγορίες σύνθετης δραστηριότητας που είναι περιγραφικές όχι μόνο για τον κυρίαρχο αλλά και για τον συνυπάρχοντα μηχανισμό δραστηριότητας που παρέχει σημαντική συνεισφορά στο παρατηρούμενο γαλαξιακό φάσμα. Τέλος, εφαρμόζουμε το διαγνωστικό μας σε ένα δείγμα φασματοσκοπικά επιλεγμένων σύνθετων γαλαξιών για να επαληθεύσουμε την εγκυρότητα του διαχωρισμού της δραστηριότητας των σύνθετων κατηγοριών γαλαξιών. Διαπιστώνουμε ότι ο διαχωρισμός της δραστηριότητας των γαλαξιών είναι πραγματικά δυνατός.

Abstract

Classification methods for characterizing the activity of a galaxy are of high importance in observational astrophysics. Even though numerous diagnostic tools have been build over the past years, the overwhelming majority of them only concerns emission line galaxies or are specialised only in one activity class (i.e, AGN). Moreover, almost none of them is able to include all possible types of activity (active and passive) under one unified scheme. Furthermore, they fail to properly address the issue of the mixed activity classes of composite and LINER galaxies. In this work, we intent to define a diagnostic tool based on machine-learning methods considering three classes that are representative of the principal mechanisms of gas excitation: star-formation, active nucleus and excitation from hot evolved stars present in passive galaxies. We use data from the SDSS and *GALEX* All-sky surveys in order to select the training sample of the active and passive galaxies. For this purpose, we train a Random Forest algorithm that utilises four features in total. Three of them are the Equivalent Widths (EW) of the spectral lines of Hα, [NII] λ6584Å, and [OIII] λ5007Å that are found to provide excellent discriminating power for the three principal types of the activity found in a galaxy. The fourth feature is the D4000 continuum break index which is a good indicator of the average age of the stellar populations. We manage to achieve accuracy of $\sim 99\%$. Due to the high performance scores achieved on the pure activity classes and based on the predicted probabilities provided by the Random Forest we can apply this method to the mixed activity classes in order to identify the dominant source of gas excitation in a galaxy. For this reason we also increase the considered activity classes to provide refined predictions for the mixed activity classes. Therefore, besides the bona-fide activity classes of star-forming (SF), active nucleus (AGN) and passive galaxies, we add mixed activity classes that are descriptive not only about the dominant but also for the secondary excitation mechanism that manages to provide considerable contribution to the resultant galaxy spectrum. Finally, we apply our diagnostic tool on a sample of spectroscopically selected composite and LINER galaxies to verify the validity of the activity decomposition of these mixed activity galaxy classes. We find that the activity decomposition of galaxies is actually possible.

Contents

1. Introduction

1.1 Galactic activity diagnostics

One of the most challenging, and important, subject in the modern observational astrophysics, is the activity classification of galaxies. The main target of the galaxy activity diagnostics is the discrimination of galaxies into categories based on their gas excitation mechanism. Usually, the identification of the radiation source that excites the gas is done based on the observed spectrum of a galaxy.

Activity diagnostics are important for the understanding of internal process that take place in a galaxy. This kind of studies help us to understand the interaction between the electromagnetic radiation emitted by a radiation source and the gas and dust structures that are found in a galaxy as each excitation mechanism produce radiation that has different spectral energy distribution, that in general, results in distinct observed spectrum. Furthermore, the development of activity diagnostic methods can give us information about the population of galaxies based on their activity (demographic surveys). More specifically, AGN demographic surveys aim to catalog active black holes that can aid our studies towards the understanding of the dust and gas accretion process in galactic cores as well as galactic evolution process in general.

1.2 Current diagnostic methods and their limitations

In order to define the various galaxy classes we first have to define the principal activity mechanisms that drive the galaxy activity. A typical galaxy contains gas and dust clouds, populations of stars that can be at various stages of their evolution and black hole at its core. Depending on the kind and the intensity of the interactions between these galaxy components, there are three fundamental sources of gas excitation: star-forming regions, an active nucleus and hot evolved stars. Obviously, as an individual galaxy can be observed at different stages of its evolution, all these sources can be simultaneously contribute to its total observed spectrum. In addition, the spectrum of a galaxy can not be uniquely attributed to a specific mechanism of excitation. For example, a population of post Asymptotic Giant Branch (post-AGB) stars in an old galaxy can mimic an active one (Stasińska et al., 2008). This complex nature of galaxies makes the process of their activity classification difficult and confusing.

Many attempts have been made and numerous diagnostic tools have been built in an attempt to solve this high significance but complex problem. Most of them are based on emission spectrum (i.e., Balmer hydrogen lines and forbidden emission lines, infrared and optical colors) in order to identify the predominant source of radiation that drives the observed galaxy spectrum and thus characterize its activity class.

One the most successful and widely used diagnostic is the Baldwin, Phillips, and Terlevich, 1981 (or hereafter BPT) diagram. The BPT is a 2-dimensional diagram that utilizes the ratios of the first two Balmer lines of hydrogen (Hα and Hβ) and two forbidden emission lines ([OIII] λ5007Å and [NII]λ6584Å). Depending on the position on the diagram of [OIII] λ5007Å/Hβ against [NII]

λ6584Å/Hα the galaxy is characterized as Active Galactic Nucleus (AGN), Star-Forming (SF), Transition Object (TO or composite galaxy) or LINER (Low-Ionization Nuclear Emission-line Region; Heckman, 1980). In addition to this diagram, diagrams involving the [SII] λ6717, 6731Å/Hα and and [OI] λ6300Å/Hα ratios are also considered (Figure 1.1).

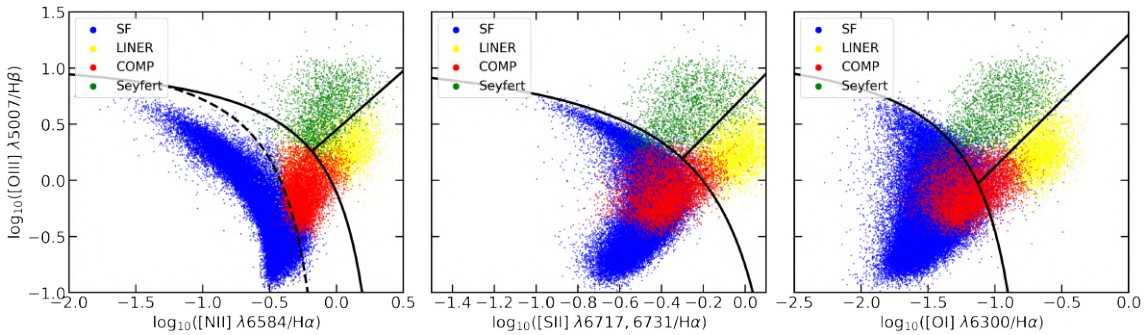


Figure 1.1: Left: standard BPT diagram, middle [OIII] λ5007Å/Hβ against [SII] λ6717,6731Å/Hα and right: [OIII] λ5007/Hβ against [OI] λ6300Å/Hα for the SDSS sample of galaxies.

Another galaxy diagnostic in the optical spectrum is that can be considered as a modified BPT but with a broader application potential is the one defined by Cid Fernandes et al., 2010 which uses optical emission lines and the Hα Equivalent Width to classify galaxies. This diagram has the advantage that galactic diagnostic methods are not limited on optical spectra. Other popular diagnostics that are based on infrared photometry are described for example in the works of Donley et al., 2012, Mateos et al., 2012 and Assef et al., 2013 which define selection criteria for AGN galaxies.

The aforementioned diagnostic tools are efficient in classifying galaxies on the categories that they have been designed to work. However, many of them are focused only in one class of galaxies (e.g, AGN) or fail to incorporate all the possible types of galaxy activity under one diagnostic scheme. For instance, the class of passive galaxies is absent from almost all classification models. Even though there are some methods for selecting passive galaxies, they are often very limited in terms of the discriminating features required in order to implemented. One additional issue arises form the fact that some galaxies may have multiple sources of ionizing radiation (e.g., star-formation and AGN, young and old stellar populations). These are the classes of composite and LINER galaxies. So far, for these mixed activity classes no diagnostic method can address the nature of their dominant gas excitation mechanism. Although these classes are recognised as having mixed activity, the continuous transition from one class to another combined with degeneracies related to metallicity, intensity of the radiation field, and shape of Hα ionizing spectrum differences in the most commonly observed and used diagnostic features in the optical spectrum are almost nonexistent and therefore the classification of these galaxies into sub-classes based on dominant activity mechanism is impossible. These two issues are normally solved with the use of Spectral Energy Distribution (SED) fitting. This method of classifying galaxies can be very effective in both tasks of selecting passive galaxies and identifying the dominant source of ionizing radiation in mixed galaxy classes but its major disadvantage is that it requires large volumes of homogeneous photometric data limiting dramatically its applicability.

1.3 A new approach in galaxy activity classification

Today, recent advances in computational methods allow us to move beyond the standard data analysis methods and use tools that just a few years ago were considered unimaginable. In addition thanks to the advent of the era of all-sky surveys, scientists have observations for million of galaxies at their disposal. These surveys are preformed at various regions of the electromagnetic spectrum. Some of of the most notable and extensively used include: the Two Micron All-Sky Survey (2MASS;

Skrutskie et al., 2006) and Wide-field Infrared Survey Explorer (WISE; Wright et al., 2010) both in infrared, the SDSS (Sloan Digital Sky Survey; York et al., 2000) in optical and the Galaxy Evolution Explorer (GALEX; Martin et al., 2005) in ultra-violet. Furthermore, the recent technological advances in computational power and the development of machine-learning algorithms allows us to analyze quickly and efficiently large volumes of data from catalogues that contain multi-wavelength data for millions of objects.

In general, machine-learning algorithms can be implemented for solving classification and regression problems in a very short amount of time, while at the same time making use of millions of data. They are especially useful in cases where the problem we intent to solve are multi-dimensional and of high complexity. Also, they are characterized by their efficiency in optimization problems. Machine-leaning algorithms can be separated into three main categories: supervised, unsupervised and semi-supervised depending on the data we use in the training process of these algorithms. In all cases we have to provide the algorithm with a number of features, i.e., different types of measurements that describe the properties of the objects we try to classify. An algorithm is characterized as supervised, if we also give the corresponding label (class) of every data point that the algorithm is going to be trained on. The goal in this case is to train it on known examples, and then applying it on different data, but similar to the ones used for its training. Supervised algorithms are mainly used in classification as well as in regression problems. Another kind of machine-learning algorithms includes the unsupervised ones. They are trained on data that does not need to have a label (class) assigned to them. Their most notable applications include pattern recognition, clustering and anomaly detection.

The capabilities of these algorithms can not be left unnoticed. They have already been applied in numerous problems across many scientific fields with extraordinary results. In particular, the use of machine-learning algorithms in astrophysics is not new. They have already been applied to many problems with high success delivering results that would have been impossible otherwise. Some examples of previous works that have used machine-learning in their analysis to solve galaxy classification problems include classification based on BPT diagram Stampoulis et al., 2019, identification of AGN properties Pennock et al., 2021 and galaxy morphology classification Domínguez Sánchez et al., 2018.

In a previous paragraph we introduced the three principal mechanisms of gas excitation that can be found in a galaxy. These correspond to the three major galaxy activity classes of star-forming (SF) or HII regions, AGN, and passive. To begin with, star-forming galaxies are rich in dust and gas and they are producing young stars. As a consequence, the emission of these galaxies is driven by blue hot massive stars. More specifically, inside an interstellar cloud there are parts of gas that are collapsing to form new stars. After the formation of stars, the residual gas forms shells around them. These shells contain gas and dust that are heated up by the UV radiation of the newly formed massive hot stars to produce strong forbidden and hydrogen emission lines. The HII regions are scattered across the galaxy disk.

The other activity class of galaxies is the class of AGN. The circumnuclear dust around a supermassive black hole located at the center of the galaxy is heated up producing extreme UV radiation that ionises the gas and dust of the torus (an area of low density and temperature located outside the accretion disk) producing forbidden and hydrogen emission lines in the visible spectrum. In general, the forbidden emission line of doubly ionised oxygen ([OIII] λ5007Å) has higher flux in an AGN environment than in an HII region. This results form the fact that the UV radiation produced form the accretion disk around the galaxy nucleus is harder when compared to that one produced by massive stars in HII regions.

Finally, we have the class of passive galaxies. These are old galaxies populated mainly by evolved stellar populations. The spectrum of this type of galaxy is characterized mainly by absorption lines. A passive galaxy has exhausted almost all of its gas and dust reservoirs. Although, in this type of galaxies we find mainly old stellar populations some of them may appear as being active (i.e, having week emission lines) due to excitation from hot evolved stellar populations.

Until now, we have described the activity of galaxies that are characterized only by a single gas excitation mechanism. Composite and LINER are two other classes of galaxies that are typically

included in the classification models as separate classes. The identification of the activity source of a composite galaxy can be complicated. These galaxies are found at the interface between star-forming and AGN galaxies in the BPT diagram and therefore can harbour more that one type of activity simultaneously. The dominant gas excitation mechanism can be either star-formation or an active nucleus. Therefore, most of the times the activity is a combination of star-formation with an active nucleus or hot evolved stars. In addition, recent studies suggest that not all composite galaxies have an active nucleus, instead the additional source of ionization can be from populations of hot evolved stars (Byler et al., 2019;Byler et al., 2017).

The mechanism that drives the activity of LINER galaxies is more difficult to be interpreted. For several years the it was thought that their activity was the result of a low luminosity active nucleus (LLAGN). Recently, some studies support the idea that the activity can be also attributed to post-AGB stars ((Binette et al., 1994;Stasińska et al., 2008;Papaderos et al., 2013). So far, all the available diagnostic methods recognise that these two classes is a result of mixed activities but fail to give any detail about the characterization of the true underlying excitation mechanism.

Although some of these methods have been successful, most of them have limited applicability, are complicated in their implementation or fail to include all the activity classes. In this work, we intent to define a new machine-learning diagnostic tool that utilises four features and is capable of discriminating galaxies into three classes that are representative of the three principal excitation mechanisms, star-formation, AGN and emission from hot evolved stars. For this purpose we use a Random Forest classifier. We choose as discriminating features the Equivalent Widths of [OIII] λ5007Å, [NII] λ6584Å, Hα, and the D4000 continuum break index (Balogh et al., 1999). The choice of using the Equivalent Widths instead of the actual flux of the spectral lines is that it allows us to include the class of passive galaxies under one unified classification scheme. In addition, the D4000 index we expect to break the degeneracy between the emission from hot evolved stellar populations that is often mistaken as emission from active galaxies. This degeneracy is particularly prominent in the case of the mixed activity classes (i.e, composite galaxies). Finally, we aim to identify the dominant mechanism of gas excitation in a galaxy as well as to identify the combination of the gas excitation mechanisms that coexists in galaxies with mixed activity classes.

2. Data sample

2.1 Data acquisition

For the needs of this project we have to combine data from two All-sky surveys. Firstly, we use the MPA-JHU (Kauffmann et al., 2003;Brinchmann et al., 2004;Tremonti et al., 2004) catalog from the DR8 data release from the SDSS all-sky survey. From there, we cross-match the *galSpecInfo*, *galSpecIndx* and *galSpecLine* catalogs. From this catalog we are interested on the EW of the lines Balmer lines of hydrogen, forbidden emission-lines of oxygen, nitrogen and sulfur as well as the D4000 continuum break of the galaxies.

In particular, we use the *galSpecInfo* to acquire information on the positions (coordinates) of the galaxies on the sky, as well as, information regarding the reliability of the measurements. The *galSpecLine* catalog provides information of the major emission lines from the SDSS spectra, after the removal of the stellar component. In particular, from the *galSpecLine* catalog we are interested in the following columns: `h_alpha_eqw`, `oiii_5007_eqw`, and `nii_6584_eqw`. These EWs refer to the Balmer line of Hα, the doubly ionised forbidden line of oxygen ([OIII] λ5007Å) and the simply ionised forbidden line of nitrogen ([NII] λ6584Å) respectively. All the EWs have been calculated from the continuum-subtracted spectrum and negative values describe emission. From the *galSpecIndx* catalog we are only interested in the `D4000_N` which has been estimated based on the definition of Balogh et al., 1999.

In order to identify a sample of passive galaxies we also use ultra-violet photometry from the *GALEX* survey. For this reason, we cross-match the SDSS sample discussed above with the *GALEX*–SDSS–*WISE* Legacy Catalog (GSWLC) from the work of Salim et al., 2016 using 1 arcsecond search radius. From the GSWLC catalog we use the `NUV` column.

After gathering all available data for the features of interest for all objects the subsequent catalog contains 206476 galaxies in total. We apply spectrum quality cuts to the whole newly composed catalog. We begin by requiring $S/N > 3$ on the continuum around Hγ spectral line. This ensures that the selected galaxies have reliable optical spectrum observations in the blue part of the spectrum. Other observables that we use as discriminating features are the EW values of Hα, [OIII] λ5007Å, and [NII] λ6584Å. For these observables we require that the continuum of the spectral line corresponding to each of the selected lines to have $S/N > 5$. These choices of cleaning the sample based on the optical spectrum (continuum) of the emission line and not the EW value itself ensures that: (a) we have an unbiased set of good quality spectra since they are not selected on the value of the features we are interested in, (b) we do not discriminate against passive galaxies which do not show emission lines. We also remove galaxies that the D4000 values is set to 0, as we discovered from inspection of the spectra that this is a result of a incorrect measurement due to the fact that some spectra do not fully cover the wavelength range in the blue part of the spectrum where the D4000 is located. Finally, the emission line measurements that are included in the SDSS catalogs are obtained by the application of a pipeline. For each object in the MPA-JHU, if the output values of objects for which the pipeline did not return reliable measurements are flagged as `RELIABLE`=0 in the SDSS catalog, these objects are removed from our analysis. The subsequent sample of galaxies following the quality cuts has 193649

galaxies.

2.2 Multi-dimensional emission-line classification of active galaxies

For the definition of our new diagnostic we choose to train a supervised machine-learning algorithm. This means that we must find the true labels of the galaxies that we will introduce to the training of the diagnostic. To find the true labels (classification) for our sample of galaxies, we use the diagnostic that was defined in the work of Stampoulis et al., 2019. In that work, a machine-learning algorithm was used for the definition of a 4-dimensional diagnostic that was based on the four emission-line ratios of log([NII]/Hα), log([SII]/Hα), log([OI]/Hα) and log([OIII]/Hβ). The model was constructed by fitting multivariate Gaussian distributions in the 4-dimensional emission-line ratio space. For each galaxy the class assignment is done based on its location in the 4-dimensional emission-line space. In Figure 2.1 we can see a 3-dimensional projection of the 4-dimensional space with the location of each class. This approach has the advantage that it considers all four features of interest simultaneously instead of their 2-dimensional projection as in the diagnostic of Kewley et al., 2006. This way we maximize the reliability of the classification while minimizing the contradictory classifications.

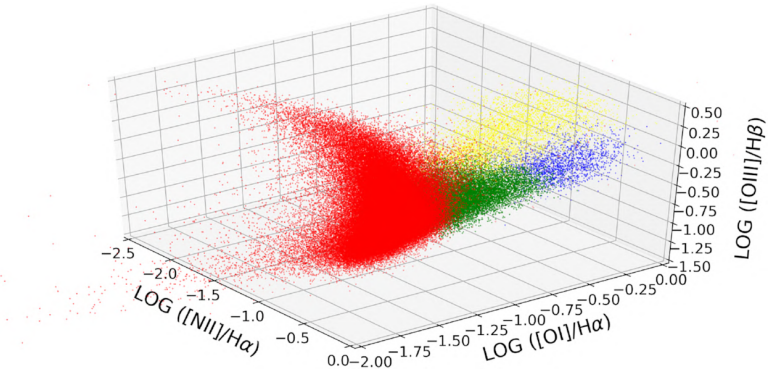


Figure 2.1: Plot of the 3-dimensional projection of the 4-dimensional emission-line space. The galaxies are color-coded based on their SoDDA classification: star-forming galaxies are marked with red, Seyferts with yellow, LINERs with blue and composites with green. Figure 7 of Stampoulis et al., 2019.

In our analysis we will use their Soft Data-driven Analysis (SoDDA) as it is based on the class with the highest classification probability. The classifications that are given as an output by this classifier characterize the galaxies into four types of active galaxies: star-forming, AGN, LINER and composite. As the diagnostic of Stampoulis et al., 2019 can be applied reliably only to galaxies that have good quality measurements on their emission line fluxes used for its definition, we apply it on our sample to obtain the classifications of the galaxies after the criteria necessary for the application have been met.

2.3 Classification of passive galaxies

After defining a sample of active galaxies, we seek to find passive galaxies. These are galaxies that must not show any clear evidence of star-forming or AGN activity. For this reason, the criteria we set in order to classify these galaxies as passive are based on the galaxy color-magnitude diagram (CMD) Bell et al., 2004. The CMD diagram is a plot of the color of a galaxy, e.g, $g-r$ or $u-r$, against its absolute magnitude in e.g., SDSS r filter, M_r. One such diagram is the $u-r$ against the M_r. The bottom area of this diagram is populated by galaxies with bluer $u-r$ colors. This area is usually referred to as the blue cloud and is mainly populated by star-forming galaxies. As we move to the top of the CMD diagram to redder $u-r$ colors we find the red sequence. This part of the diagram is

populated by red galaxies. As mentioned in the work of Haines, Gargiulo, and Merluzzi, 2008, the bimodality between the red sequence and the blue cloud galaxies can be used for a reliable selection of a passive galaxy sample.

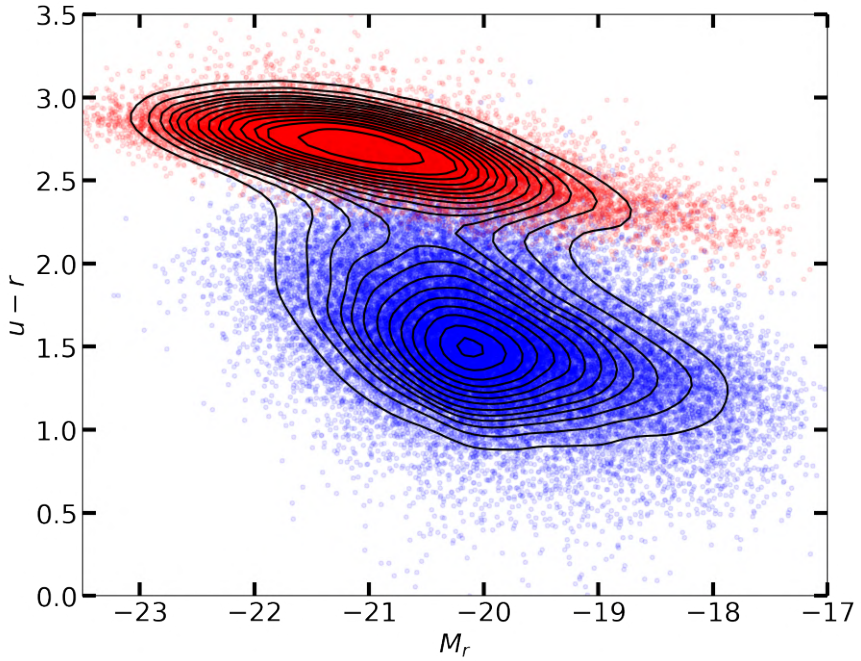


Figure 2.2: CMD diagram of $u-r$ against M_r. The red dots represent the red sequence galaxies. The blue dots form the blue cloud. Sample of galaxies taken from SDSS.

On that work, the authors provide two selection criteria for obtaining a sample of passive galaxies: (1) $u-r > 2.291 - 0.1191 \times (M_r + 20) - 0.181$ and (2) $NUV - r > 5.393 - 0.1782 \times (M_r + 20) - 0.370$. After choosing the selection criterion (1) we find that it is not enough to separate red sequence and blue cloud galaxies adequately. For example some star-forming galaxies have significant dust obscuration, making them to appear much redder than they actually are. As also mentioned by Haines, Gargiulo, and Merluzzi, 2008 all passive galaxies are red but not all red galaxies are necessarily passive.

In fact, we find that there is a significant fraction of spectroscopically classified star-forming and AGN galaxies that satisfy the selection criterion (1). This results in a non negligible contamination of the passive galaxy sample. Since we need a very clean sample of passive galaxies for properly training our diagnostic, to overcome this problem, we choose selection criterion (2) which ensures the purity of our passive galaxy sample. In addition to the above criterion we also remove of any galaxy that has been spectroscopically classified as star-forming or AGN. In other words, although in our sample of passive galaxies we have included galaxies that exhibit characteristics of an active one, a passive galaxy that has strong emission-lines and it is classified as SF or AGN we adopt the latter classification. From the selected sample of passive galaxies we choose not to remove any emission-line object that has been spectroscopicallly classified as LINER or composite since these may be associated with evolved stellar populations.

In the case of LINER galaxies, even though it is considered that the gas ionization mechanism is an active nucleus, there are evidence that their emission can be attributed to hot evolved stellar populations (post-AGB stars) e.g., Singh et al., 2013. In the case of composite galaxies it is generally considered that the origin of their activity is an active nucleus along with a star-formation component. However, as mentioned in the work of Byler et al., 2019 it is also possible that the activity of these galaxies can be attributed to weak residual star-formation aided by ionization from hot evolved stellar populations. Even though these two sub-populations of LINER and composite galaxies have emission

lines that otherwise would have characterized them as active galaxies the main ionization mechanism results from old stellar populations. By including objects with weak emission lines in the training sample we provide the diagnostic with information about old stellar populations as a feature of passive galaxies. In Figure 2.3 we present the sample of passive galaxies that was selected with the criteria mentioned above in a $(g-r)$ against M_r CMD as an additional verification of the efficacy of the followed selection method.

Finally, the u and r optical SDSS colors of the galaxies have been corrected for Galactic dust extinction based on the Cardelli, Clayton, and Mathis, 1989 extinction law with $R_V = 3.1$ and $E(B-V)$ values that are from the dust maps of Schlegel, Finkbeiner, and Davis, 1998.

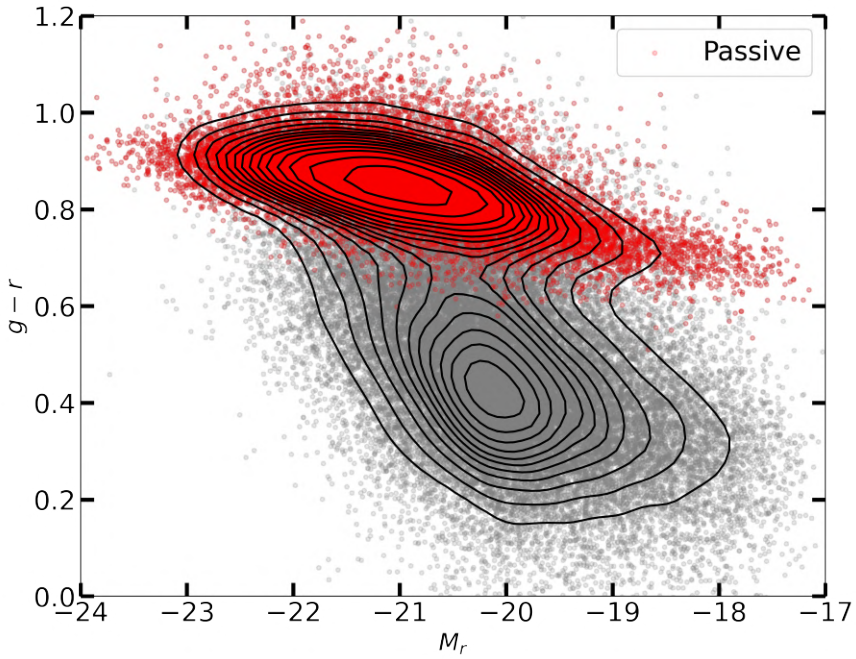


Figure 2.3: CMD diagram of $g-r$ against M_r. The red dots represent the selected sample of passive galaxies. The grey dots are galaxies from all classes of the sample. We observe that the galaxies we selected as passive actually belong to the red sequence. The contours represent the density of the objects.

2.4 Data processing and final sample

After we have obtained the data from the catalogues the galaxies were classified into passive (dormant, hot evolved stars) and active (SF or AGN) galaxies. The next step of the procedure of data processing is to select galaxies that exhibit good quality on their observed features of interest. This is a crucial step towards defining a machine-learning diagnostic because by including low quality data in its training will result in high uncertainty (high mixing between the different classes) during the training but also poor predictions when applied to similar but unknown data. In section 2.2 and 2.3, we followed two different ways to select passive and active galaxies as their nature is fundamentally different and so far no diagnostic can provide a unified classification scheme. Starting with passive galaxies, in the process of selecting passive galaxies we were based on the CMD of the color $NUV - r$ against the r-absolute magnitude (M_r). To ensure the reliability of the classification we set a signal-to-noise (S/N) selection criterion. Any passive galaxy that has $S/N > 3$ in $NUV - r$ color is included in the final sample.

The rest of the data sample contains the active galaxies (i.e, SF, AGN, LINER and composite). These galaxies were selected after the application of the diagnostic tool of Stampoulis et al., 2019. As this classification method utilizes optical emission-lines we have to ensure that only galaxies with high quality observed spectra are in the final sample. This is achieved by setting a $S/N > 5$ selection criterion for all emission-lines that were used to classify these galaxies, namely Hα, Hβ, [OIII] (λ5007Å), [OI] (λ6300Å), [NII] (λ6584Å), [SII] (λ6717Å) and [SII] (λ6731Å).

As previously mentioned in the introduction, we are interested in defining a diagnostic that incorporates all fundamental types of galactic activity. Thus we are considering only the three main types of activity which are star-formation, active nucleus and old stellar populations. For this reason, in the final sample that will also represent the training sample we only include galaxies that has been classified as SF, AGN and passive based on the aforementioned selection process. We are eliminating all other galaxy classes (i.e, composite and LINER galaxies) never to be present in the training sample. The composition per class of the final sample that is will be used for the training of the algorithm is presented in table 2.1.

Table 2.1: The composition per galaxy class of the final sample. This sample will be later used for the training of the algorithm.

Class	Number of objects	Percentage (%)
Star-forming	36287	56.9
AGN	1435	2.3
Passive	26007	40.8
Total	63729	100.0

We acknowledge that the BPT diagram is a very efficient and well-established method for classifying active galaxies. For this reason we plot projections of our training sample on the BPT diagram in order to check the distributions of our classes based on standard methods. In Figure 2.4 we present the BPT diagrams that indicate the distribution of the training set for each one of the classes three classes, SF, AGN and passive projected on the log(([OIII] λ5007Å/Hβ) against log(([NII] λ6584Å/Hα), log(([SII] $\lambda 6717,6731$Å/Hα) and log(([OI] λ6300Å/Hα). The subset of passive galaxies that is present on these diagrams have $S/N > 3$ in all emission-lines used for the plots. Of course, due to their nature a large portion of the passive galaxy sample is absent from these plots as their extremely weak emission or nonexistent (absorption) lines produce low S/N excluding them from this classification scheme.

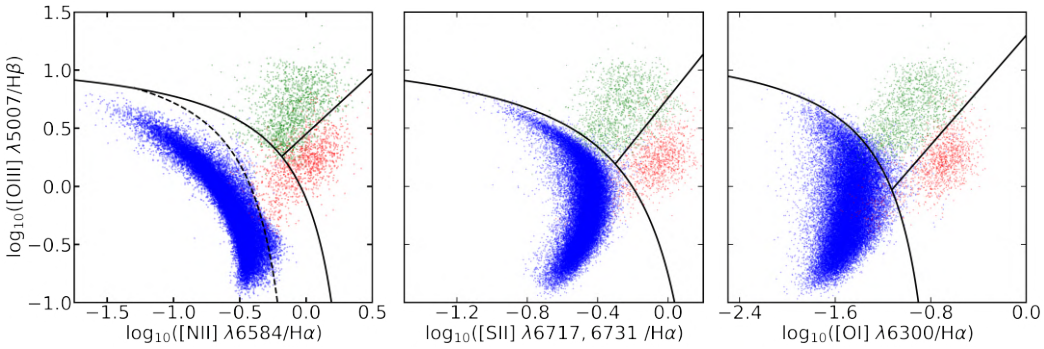


Figure 2.4: Projections of the training sample of the three galaxy classes on (a) the standard BPT diagram, (b) on the log([OIII]/Hβ) against log([NII]/Hα) and (c) on the log([OIII]/Hβ) against log([SII]/Hα). The SF are the blue dots, the AGN are the green dots and the passive are the red dots. In all three plots only a subsample of passive galaxies is presented (emission-lines of the required emission-lines for the plots $S/N > 3$).

2.5 Feature selection

There are a number of features that can be chosen to discriminate effectively between the various classes of galaxies. In the past, emission-line ratios have been used successfully for the characterization of the galaxy activity class. However, emission lines are only present for active galaxies as older galaxies have little to no dust and gas reserves to produce them. By introducing the EW of a spectral line instead of the emission-line flux itself, we intent to overcome these two problems and define a self-consistent diagnostic that can be used seamlessly for galaxies exhibiting emission as well as absorption lines. The EW of a line is defined as $EW = \int_{\lambda_1}^{\lambda_2} \frac{F_{cont}-F_{line}}{F_{cont}}\, dx$, where the F_{cont} and F_{line} represent the continuum and spectral line flux respectively. The λ_1 and λ_2 represent the wavelengths of the limits of the spectral line region. Here, negative values for the EW correspond to emission while the positive values correspond to absorption. The adoption of EW offers two main advantages that make their use in a diagnostic superior to flux or flux rations. First, EW can be available for a wider number of galaxies as the criterion for reliable EW values can be set from the quality of the continuum spectrum on either side of the line and not from the intensity of the spectral line itself which may result in bias. Also, we can include the class of passive galaxies which is a matter of high significance if we consider the fact that the lack of emission lines normally excludes them from the standard activity diagnostic methods.

Our next step is to find the minimum number of spectral lines that can identify the driving ionization mechanism in a galaxy. This will result in an efficient classifier that can accurately discriminate between classes while at the same time it increases the applicability of the classifier to a wider number of datasets.

In the quest of finding these optimal features we start by thinking both astrophysical and practical reasons for each potential feature scheme selection. From an astrophysical point of view, we are partially motivated by the physics of the emission line diagnostic methods and we consider the EW of the lines of Hα, [NII] λ6584Å, [OIII] λ5007Å, [SII] λ6717,6731Å and [OI] λ6300Å. There are physical reasoning behind this motivation as we expect the EWs of the doubly ionised oxygen will be higher in an AGN environment than in an HII region. This can be explained form the fact the UV radiation in an AGN environment is harder than that typically found in a HII region. From practical point of view, we choose strong spectral features that are generally easy to be observed and measured (e.g, Balmer lines, [OIII] λ5007Å). However, as since the EW(Hα) and the EW(Hβ) appear to be highly correlated for almost all classes of galaxies, the inclusion of one makes the use of the other redundant. Therefore it is meaningless to include both of them and we choose to use the Hα that has stronger EWs and thus it is available for a broader number of galaxies than the Hβ.

Following a similar argument we also see that the EW([SII] λ6717Å,6731Å), EW([OI] λ6300Å) show very similar behaviour as the EW(Hα) and [NII] (λ6584Å) line. This observation indicates that the inclusion of these [SII] (λ6717Å,6731Å), [OI] (λ6300Å) in our new diagnostic scheme may not offer much additional information about the differences of the principal galaxy activity classes.

Finally, in addition to these features we also want a feature that carries the information of the age of the stellar populations in a galaxy. This way, we attempt to achieve our second goal which is to identify the true nature of the mixed activity classes. This can be achieved with the use of a feature that has the properties of a stellar age indicator. After considering many possible age sensitive indicators (e.g, Hδ), most of them are not reliable as there is a degeneracy between the age of the stellar populations and metallicity. This is because most of these indicators do not scale monotonically with the age of the stellar populations. However, we find that the D4000 continuum break can be an excellent indicator for the estimation of the average age of the stellar populations in a galaxy. The rationale behind this is that the amplitude of the D4000 index is affected mainly by the massive stars of the main sequence (MS) and therefore it increases almost monotonically with the age of the stellar populations. The D4000 index is defined as the amplitude of the discontinuity in the blue part of the spectrum. Is is calculated as the ratio of the average flux in a blue to the average flux in a red narrow wavelength interval. The blue wavelength interval has range of (λ_1^{blue},λ_2^{blue}) = (3850Å, 3950Å), while the red wavelength interval has range (λ_1^{red},λ_2^{red}) = (4000Å, 4100Å). The amplitude of D4000 break is calculated as $D4000 = \frac{<F^{blue}>}{<F^{red}>}$, where $< F^{blue} > = (\lambda_2^{blue} - \lambda_1^{blue}) \int_{\lambda_1^{blue}}^{\lambda_2^{blue}} F_\nu \, d\lambda$ and $< F^{red} > = (\lambda_2^{red} - \lambda_1^{red}) \int_{\lambda_1^{red}}^{\lambda_2^{red}} F_\nu \, d\lambda$. The definition for the D4000 we have adopted here is narrower that the usual (Bruzual A., 1983). We consider the D4000 definition of Balogh et al., 1999 that has narrower wavelength intervals allowing us to include more galaxies while at the same time it is less reddening sensitive.

Concluding, we found that using the EW of the Hα, [OIII] λ5007Å and [NII] λ6584Å lines along with the D4000 continuum break index, we can define a diagnostic tool that fulfills all proposed criteria. The distributions of each feature per class is presented in Figure 2.5.

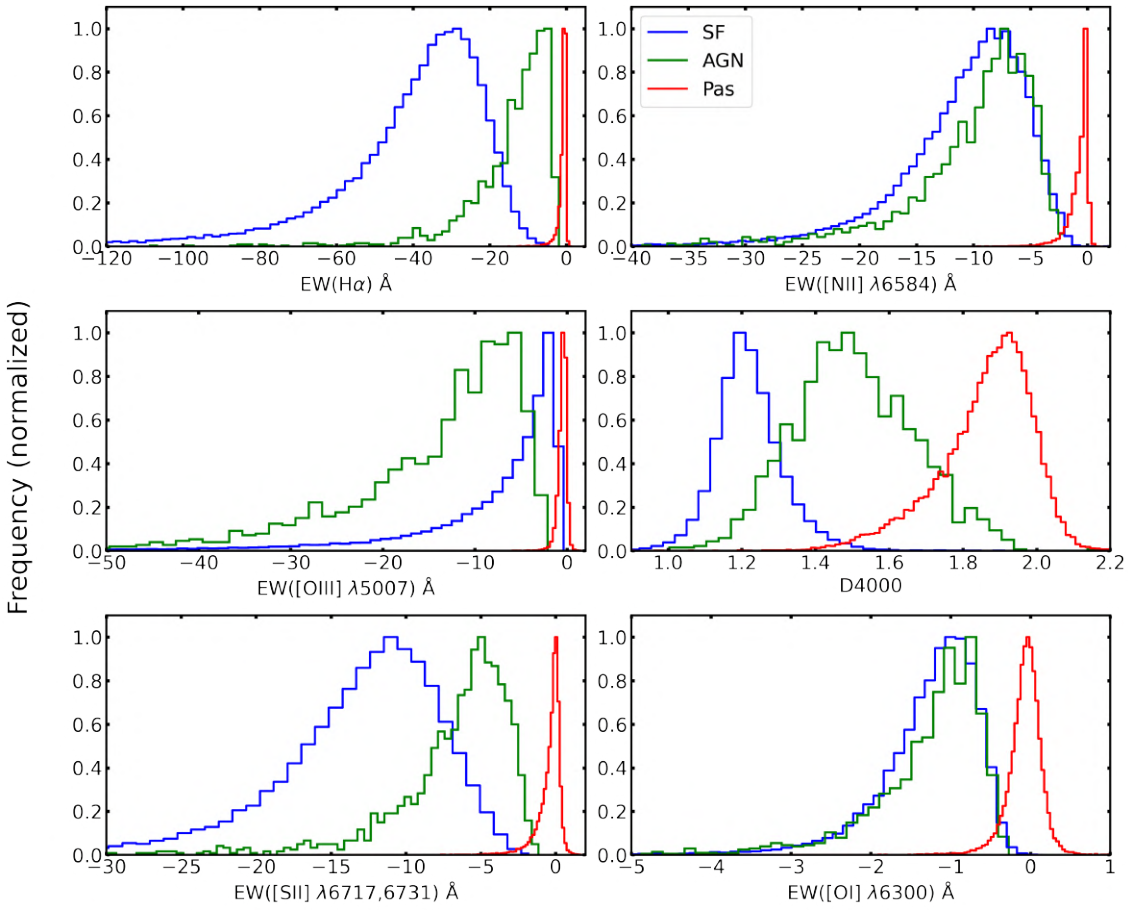


Figure 2.5: Distributions of six potential features for the three principal activity classes of star-forming (SF), AGN and passive (Pas) galaxies, top left: EW(Hα), top right: EW([NII] λ6584), middle left: EW([OIII] λ5007Å), middle right: D4000, bottom left: EW([SII] λ6717,6731Å) and bottom right: EW([OI] λ6300Å) for each of the three activity classes star-forming, AGN and passive. These are the EW of the corresponding emission-lines that are commonly used in galactic activity classification models.

3. Random Forest classifier

3.1 The Random Forest classifier

There are numerous machine-learning algorithms that one can choose to implement in classification problems. Firstly, we have to choose if our algorithm will be supervised on not. By selecting one over the other there are certain advantages and disadvantages. For example if we choose a supervised algorithm we have to find the true labels (classifications) for the data we are going to use. This can potentially reduce the total number of objects available for the training which can lead to low performance. On the other hand, unsupervised algorithms are mostly suitable in finding structures and correlations in a sample of data. As in this work we are interested in classification of galaxies, i.e defining a diagnostic tool, the most appropriate choice is a supervised algorithm.

In this project the problem we are trying to solve is a 4-dimensional classification problem. As mentioned earlier, the complexity of the problem requires the use of a flexible algorithm that can discriminate efficiently between the galaxy activity classes. The Random Forest classifier is a supervised classification algorithm that has been extensively used for the development of many diagnostic tools in many different fields. The reason for this is because it offers great flexibility as there are many parameters that can be tweaked and tuned to fit the needs of each individual problem. Also, this algorithm is known for delivering robust results as its training is not generally affected by outliers. Another benefit of the Random Forest is that its operation is simple and intuitive.

The Random Forest is an ensemble classifier. This type of classifiers combine a number of small models into a single unified model. One advantage is that the final model performs significantly better than the individual models that where used to built it. In other words, this method allows us to use many classifiers in parallel to increase their performance. An additional benefit, is that we can avoid overfitting our data, which occurs when the model has learned the training data too well. In this case the classier achieves almost perfect scores on the subset of data that was trained on. When this happens the performance of the model drops sharply when it tries to classify new data which are not similar to the ones that were included in its training.

The Random Forest model as an ensemble classifier consisting of many individual decision tree classifiers. In order to understand how the training process its done we have to look at the decision tree. A decision tree has a specific structure; it starts with the root node, followed by the inner nodes and it ends with the leaf nodes. The training process starts from the root node. The data that enter there can contain all or a randomly selected portion of the available training data. Afterwards, with the use of the features that we have selected, the decision tree splits the data making progressively purer nodes. That means that each subsequent node contains mostly objects that share similar characteristics based on the selected features, or in other words, they belong to the same class. The splitting of the data at each node is done based on the impurity value of the subsequent node. If the impurity of the produced node is lower than the parent node then the splitting is performed. The are two ways to measure impurity; the Gini metric and the entropy. Both of them describe the information gain after each node splitting. The Gini impurity is defined as the probability of misclassifying an object. Lowering the value of impurity at each node results in a classifier with low instances of misclassifications. The

node splitting process stops when further splitting of the data into new nodes does not lower the value of impurity. The final nodes are called leaf nodes and are the purer nodes in a decision tree. By making an ensemble of many of these decision trees we then build a Random Forest classifier. It is called random as the data that are included for the training of each individual decision tree is selected randomly, and because the features used for the data splitting at each node are selected randomly.

Along with the classification labels the Random Forest also calculates the probability of the classified object to belong in each one of the classes considered in the classification problem. Each tree in the end of the classification process assigns a class to a given object. The final adopted class is based on the summary of the votes from all the tress, for each object. The class that has received the majority of the votes is the adopted class for each object. Since all the trees in the ensemble vote for each object individually, the probability that the object under investigation belongs to each one of the classes is defined as the fraction of votes for this specific class to the total number of votes (i.e., the number of tress).

3.2 Implementation

The implementation of the Random Forest algorithm we adopt is the `RandomForestClassifier()` in the `scikit-learn` Python 3 package, version 1.1.1 (Pedregosa et al., 2012). We provide the algorithm with four features, namely [OIII] λ5007Å, [NII] λ6584Å, Hα, the D4000 continuum break and the activity class of each object. Based on these features it is trained to classify galaxies based on the three main types of activity, star-forming (SF), AGN and passive. The performance of this algorithm is mainly driven by the following hyperparameters: `max_depth`, `max_leaf_nodes`, `max_samples`, `min_samples_leaf`, `min_samples_split` and `n_estimators`. A more detailed description for them is given in table 3.1. These are the hyperparameters that have the higher impact when we try adapt the Random Forest to a specific classification problem and hence on the resulting performance.

Table 3.1: Description of the all hyperparameters with the higher impact on the Random Forest performance.

Hyperparameter	Description
`n_estimators`	Total number of trees in the Random Forest ensemble
`max_leaf_nodes`	Maximum number of the leaf nodes
`max_samples`	The number of samples chosen to train each tree
`min_samples_leaf`	The minimum number of objects that exist in leaf node
`min_samples_split`	Minimum number of objects needed for an internal node to split
`max_depth`	The total number of splits each tree is allowed to make.
`bootstrap`	If set to `True`, the trees will be trained on a randomly selected subsample of the original data.
`class_weight`	The inverse of the appearance frequency of each class in the training sample.
`criterion`	Function that measures the quality of each split during the node creation. Has two options: `entropy` and `Gini`

From further investigation we verify that the rest of the hyperparameters do not affect the performance of the and they are left to their default values as imported with the `RandomForestClassifier()`.

After the selection of the hyperparameters the next stage in the definition of the diagnostic is to train it. For this stage we use the data that were selected under the specific criteria mentioned in section 2. The whole data sample is spit into two subsets: training and test set with a 70%-30% ratio respectively. The split we perform is a stratified one, which ensures that each one of the subsets of data contain the same fraction from each individual class. Before this splitting step the data undergo a random shuffling to ensure homogeneity of the data in the two subsets. The training set, which has the majority (70%) will be used for the training processes of the algorithm. These are the data that the trees of the Random Forest are going to be built from. The test subset, which contains the remaining 30% of the sample, is going to be used exclusively for the evaluation of its performance. This guarantees that the algorithm performs well to similar but unseen data and thus its use can be generalized and applied to detests other than the one used for its training.

3.3 Performance Metrics

For the evaluation of the performance of our diagnostic tool we adopt standard performance metrics. To begin with, we want to know not only the fraction of correct predictions but we are also particularly interested in the misclassified instances. This kind of analysis can reveal not only how successful the classifier is but also its weaknesses and limitations which are manifested as mixing in the predicted labels between the different classes. To visualize the performance we calculate the confusion matrix based on a sample of objects that we already have he true classifications (test set). The next step, is to implement our diagnostic on this particular subset of data so that we find the predicted class. Then, by comparing the class predictions and the corresponding true class labels we make a matrix that in its rows we have the true class labels and in columns we have the predicted class labels. This leads to the conclusion that a confusion matrix that has elements only in its primary diagonal ($y = -x$) describes a perfect classifier. If the classifier has some misclassified instances, then the off-diagonal elements will be non zero giving us a detail picture about the misclassified objects. In Figure 3.1 an example confusion matrix is presented concerning a classification problem with three classes.

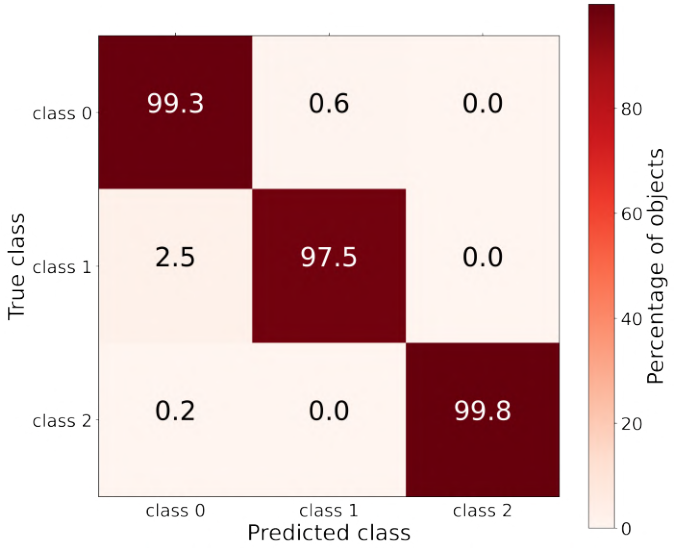


Figure 3.1: Example confusion matrix for a Random Forest model of a 3-class classification problem. The color bar represents the percentage of objects in each box calculated based on the total number of objects per true class.

In addition to the confusion matrix there also other useful performance metrics that can give us a more quantitative analysis of the overall performance but also performance for each class separately. The most common one is the accuracy score. Accuracy is the fraction of the correct predictions to the total number of predictions. However, the accuracy alone can be extremely misleading, especially in classification problems with high class imbalance. For this reason along with the accuracy we calculate the recall score which is a measure of completeness, how many objects of the same class have been successfully retrieved. Another metric is the precision score which is measure of contamination of our predictions. Contamination is an estimation of the number of objects that have been predicted to belong to a class but their actual true class was different. Finally, an additional metric we can use is the F_1-score. It is defined as the harmonic mean of the recall and precision scores. A detailed description of these metrics is presented on table 3.2, where 3.2 the performance scores are described for the case of a binary classification. These metrics can easily be generalized for multiclass classification problems.

One additional method of estimating the discriminating efficiency of a classifier is the Receiver Operating Characteristic or ROC curve. By plotting the sensitivity (True Positive Rate or recall)

Table 3.2: Description and definition of every performance metric that is used for the evaluation of the performance of the diagnostic.

	Description	Equation
Term		
True Positive (TP)	A object that has been correctly classified based on its true class label.	-
True Negative (TN)	An object is correctly classified not to belong to the class.	-
False Positive (FP)	An object that is falsely classified to belong to a class.	-
False Negative (FN)	An object that is falsely classified not to belong to the class.	-
Performance metric		
Accuracy	The fraction of the correct classifications to the total predictions that the classifier made.	$\frac{TP+TN}{TP+TN+FN+FP}$
Balanced accuracy	The average of the recall scores from each class.	$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)$
Precision	The number of objects correctly predicted to belong to a class divided by the total objects that the were classified to belong to that class.	$\frac{TP}{TP+FP}$
Recall	The objects correctly classified to belong to a class divided by all the objects that truly belong to that particular class.	$\frac{TP}{TP+FN}$
F_1-score	The calculated harmonic mean of reacll and precision.	$\frac{2TP}{2TP+FP+FN}$
Specificity	The fraction of negative examples that have been predicted as negative	$\frac{TN}{TN+FP}$

against the 1-specificity (False Positive Rate) we can inspect how well the diagnostic can discriminate one class against all the other ones. This is true because sensitivity is a measure of how well the diagnostic can detect positive examples, or in other worlds, it describes the ability of the diagnostic to predict the true positive examples for each class. The other metric of this plot, specificity, is a measure of the true negative examples that were correctly identified by the classifier. Thus, based on the above definitions, plotting the ROC curve we can observe that for a perfect classifier the area under the curve will be maximum (or 1), while for a classifier that does not predict better than random the curve will be a diagonal line with slope of one ($y = x$). An example ROC plot for a binary classification problem is presented in Figure 3.2.

The ROC curve is defined for a binary classification problems but it can be used in multiclass problems, like the one we have in this work. This is possible if we break the multiclass problem into many binary ones, in one-vs-rest fashion. A way to quantify the ROC plot is to calculate the area under the curve (AUC). A value of AUC that approaches 1 is indicative of a perfect classifier with high discriminating power.

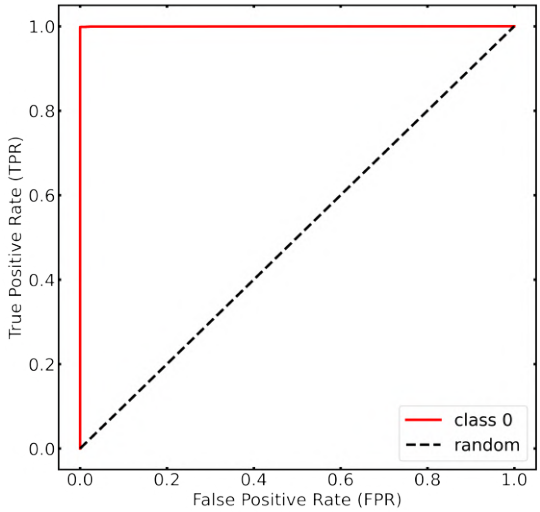


Figure 3.2: Sensitivity against 1-Specificity (ROC curve). The red line indicates the True Positive Rate agiast the False Negative Rate for a binary classification for the class 0 (consider as the positive examples) of a perfect classifier while the black dashed line describes the results for a classifier making random predictions.

3.4 Algorithm optimization

In order to make the algorithm fit the needs of our classification problem and achieve its optimal performance we have to tweak ever hyperparameter that has significant impact on the performance. These have been mentioned in detail in section 3.2. As each individual numerical hyperparameter can take values over a wide interval, it is impossible to find them manually by trial and error.

The other method to find these optimal values, is to calculate the performance on a multidimensional grid of values of these hyperparameters. Given the large number of hyperparameters and wide range of possible values in order to reduce the number of grid points and make the problem more tractable we use the validation curves. In Figure 3.3 we plot the validation curves for all significant hyperparameters that impact the performance of the classifier.

To asses the performance of the algorithm we use the k-fold cross-validation (CV) method to check the accuracy score and ultimately select the best set of hyperparameters. Cross-validation is a method where the splitting of the training and test data is performed in k times using different randomly selected subset of the data. The training set is split in k stratified set of equal size (folds). The training of the algorithm is performed k times and the scores that are reported are the average of all k scores. In every training cycle k-1 folds are are used for training and the kth fold for the testing of the performance. Also, in every training cycle the testing fold is substituted by one of the training folds. The process is repeated until the algorithm performance has been tested on all individual folds. One important benefits of using this method is that we can have an estimation of the uncertainty of the scores.

These curves describe how a chosen performance score, in our case accuracy, varies as a function of the different values that a particular hyperparameter can have. In more detail, the method we use to make the investigation about the ranges of the hyperparameters is the following: each time we choose one hyperparameter of interest we vary it in a range of its possible values while at the same time keeping all other hypreparameters in their default values. Then, we record the desired performance score which is the balanced accuracy (see 3.2) for all the selected range of its possible values. We repeat the same process for all the other hyperparameters.

The algorithm we use to perform the task of optimization is the `GridSearchCV` which is provided by the `scikit-learn`. After the inspection of the validation curves we find the optimal ranges and we use them as an input of the `GridSearchCV` algorithm to define the grid for the multidimensional

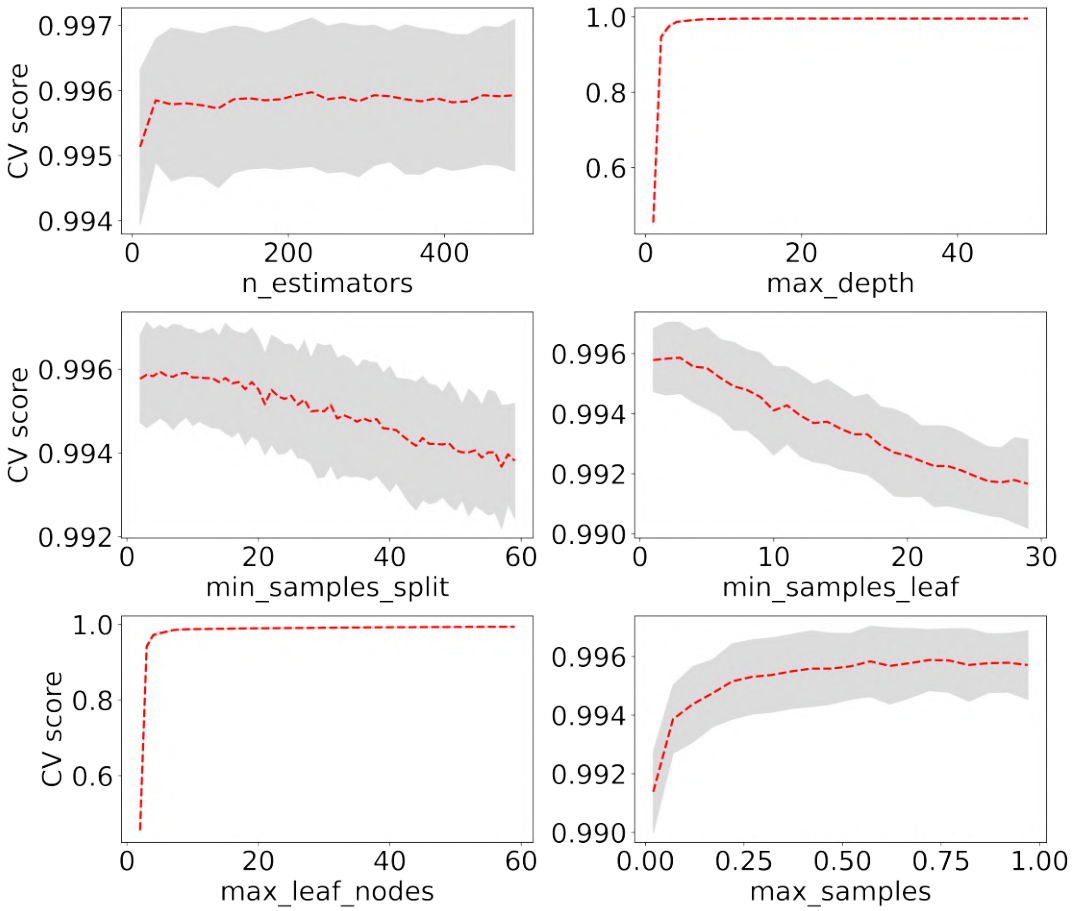


Figure 3.3: Validation curves for the six hyperparameters that have the higher impact on the Random Forest performance. The CV score represents the average balanced accuracy of the model as a function for each hyperparameter grid (red dashed line) and the error (grey shaded area) calculated using the k-fold cross-validation method (here k=10).

search. We do not need to search the values of all hyperparameters. Form the validation curves we deduce that the hyperparameters of `max_depth` and `class_weight` does not need to be optimized as every value other than the default results in overfit. Thus we leave these two parameters on their default value as initially imported by the `scikit-learn`. The ranges and the optimal hyperparameter values are presented in the table 3.3.

Table 3.3: Heperparameter search ranges and optimal values.

Parameter	Search range	Best value
`n_estimators`	100-200	160
`max_depth`	-	`'default'`
`min_samples_split`	25-40	38
`min_samples_leaf`	2-20	7
`max_leaf_nodes`	-	`'default'`
`max_samples`	0.1-1.0	1.0
`class_weight`	-	`'balanced'`
`criterion`	-	`'Gini'`

The values presented in table 3.3 are those adopted for our implementation of the Random Forest

algorithm.

3.5 Feature importance

In a section 3.1, we mentioned that the Random Forest is an ensemble of many decision trees which during their training process split the data on different classes based on the selected features. Even though all features that have been selected are important, some of them can have higher impact than others. By finding the most important features we can identify if there are any redundant features and remove them, in a effort to reduce the complexity of the model. Reducing the number of features to what is absolutely necessary, establishes an efficient classifier and increases the its applicability to a wider range of datasets. In addition, it can help us interpret the results better and gain more insight into the astrophysics of the classification problem. From this plot we can see that there is a systematic (but not very significant) decrease in the importance of the considered feature from the Feature 2 to the Feature 3 and to Feature 1. However, we do not see any feature that has significantly lower impact that the rest, therefore, all the features are required by the classifier.

A metric that is commonly used for this analysis is the feature importance. We use the Gini importance (i.e., the mean decrease of impurity) which is provided by the `scikit-learn` package. The Decision Trees use the features to create purer nodes at each consecutive split of the data. The criterion for the creation of a new node is the impurity reduction based on the gini impurity. After the training of the Random Forest we can calculate the average decrease of the impurity for each feature. Then, by taking the average over all the trees of the ensemble we find the measure of the feature importance. An example plot of feature importance for a classification problem is presented in Figure 3.4.

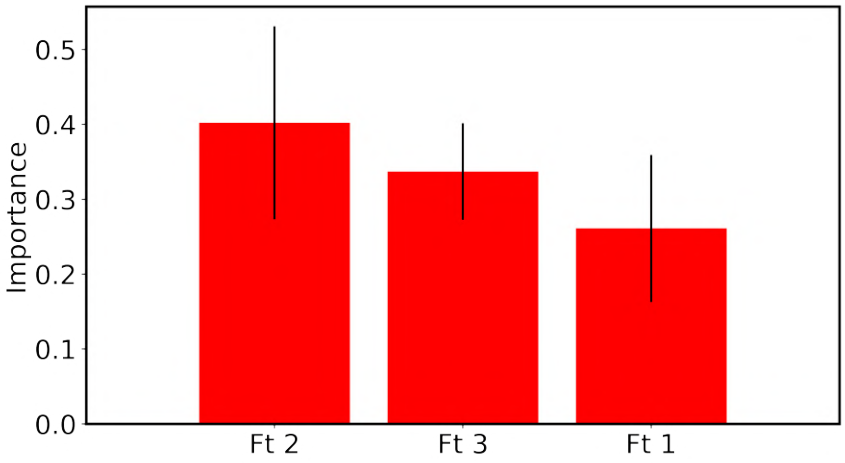


Figure 3.4: Example plot of the feature importance from a model with 3 features (Ft 1: Feature 1, Ft 2: Feature 2, and Feature 3: Ft 3). We see that the feature Ft 2 is more important in the process of splitting the data into new nodes. We also see that all the features have similar importance so there are not any redundant features. The error bars is the standard deviation of each average score.

4. Results

4.1 Performance of the three main activity classes

After the training and optimization of the algorithm we want to check the performance of the diagnostic tool on a more general case. For this reason, we utilise the test subset to perform this analysis. This is an important procedure as we can not only check how accurate the classification is but also we can inspect in detail the behaviour of the diagnostic. For example, we can identify misclassifications and the mixing between them. Furthermore, we can identify potential limitations and weaknesses of the new diagnostic tool.

We perform this analysis using the aforementioned performance metrics (section 3.3). We find that the overall balanced accuracy (see table 3.2) we achieve is 0.989 ± 0.004. This is the average balanced accuracy calculated with the method of k-fold cross-validation for k=10. The uncertainty is calculated as the standard deviation of the k-fold averages. In table 4.1,we summarize the performance scores of precision, recall and F_1-score for each galaxy class separately.

Table 4.1: Report of performance scores calculated on the test sample for each galaxy class for three different metrics.

Class	Precision	Recall	F_1-score	Galaxies
Star-forming	1.00	1.00	1.00	10800
Seyfert	0.88	0.98	0.93	464
Passive	1.00	0.99	1.00	7846

From table 4.1 we can see that the scores of all classes are nearly perfect. The high recall score of each class tell us that the classifier is able to retrieve nearly all objects of every class correctly. Based on this fact we can deduce that the classifier has high completeness. In addition, the high precision scores tell us that there are only a few instances where objects have been predicted to belong to a different class other than their true class meaning that the contamination in each class is low.

Another more detailed method to evaluate the performance of the algorithm is to plot the confusion matrix. In Figure 4.1 we present the confusion matrix for this diagnostic calculated on the objects of the test subset. By inspecting the confusion matrix we can not only identify the fraction of objects that have been correctly classified (principal diagonal elements) but also the percentage of the objects that have changed classification (off-diagonal elements). In addition, we can also find what is the preferred class of those misclassified objects giving us the information about the nature of the problem.We see that the confusion matrix is nearly diagonal. Only a very small fraction (1.2%) of the AGN are misclassified as star-forming, which is tolerable.

Finally, as one of the purposes of this work is the the definition of a diagnostic that can identify the main ionization mechanism in a galaxy, we have to ensure that the classifier is able to separate the different kinds of activity as effectively as possible. The method we use to verify that our diagnostic has high discriminating power is to plot the ROC curve. We present the ROC curves for the SF, AGN and passive galaxies in Figures 4.2, 4.3 and 4.4 respectively. We see that for all classes the

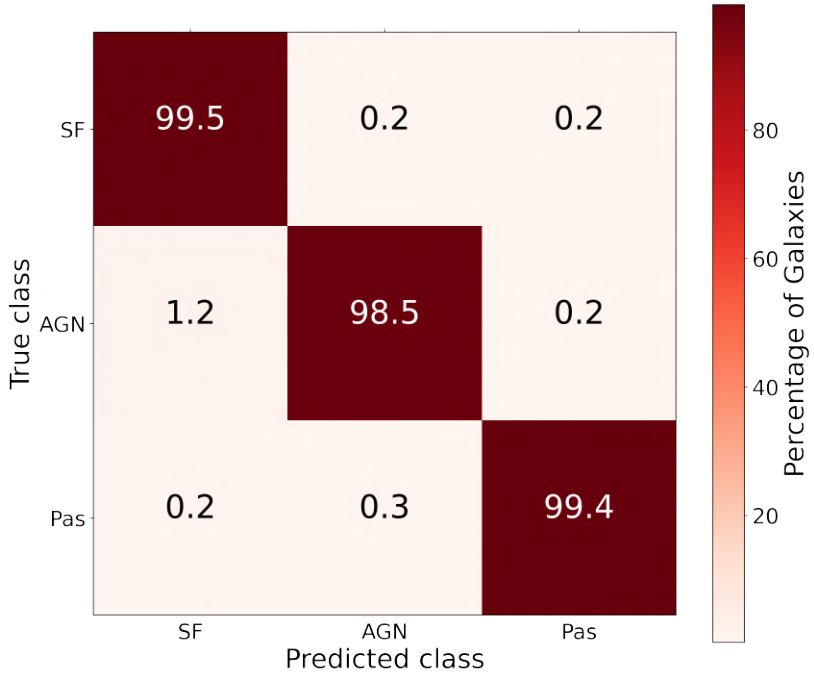


Figure 4.1: Confusion matrix calculated on the test subset of the final sample. The color and the number in each box refers to the percentage of objects calculated on the total number of true instances for each class separately.

discriminating power of the diagnostic is nearly perfect for all classes.

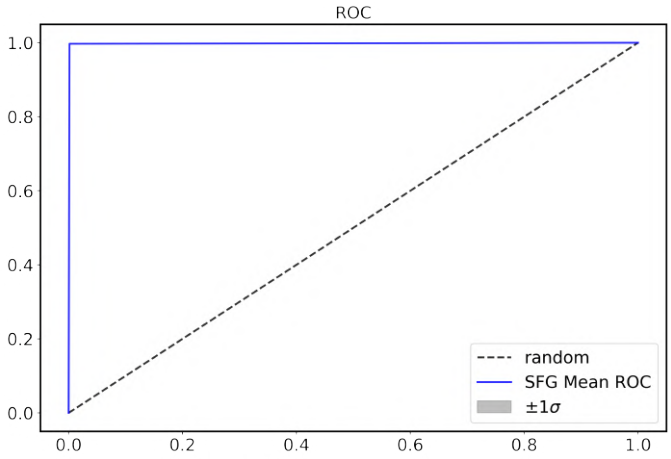


Figure 4.2: The ROC curve for the class of SF galaxies (blue solid line). The area under the curve (AUC) is above 0.99. The calculation of the mean ROC performed with the k-fold (k=10) cross-validation method. The error is the 1σ of the standard deviation of the k scores. For comparison, we plot the black dashed line which represents the ROC of a classifier that makes random predictions. Errors are too small to be shown.

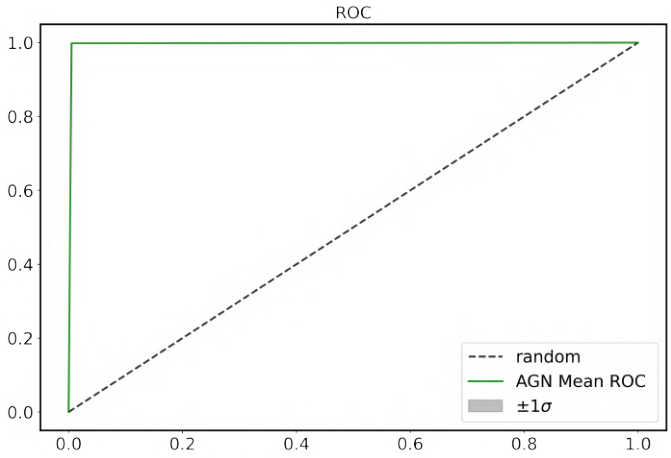


Figure 4.3: The ROC curve for the class of AGN galaxies (green solid line). The area under the curve (AUC) is above 0.99. The calculation of the mean ROC performed with the k-fold (k=10) cross-validation method. The error is the 1σ of the standard deviation of the k scores. For comparison, we plot the black dashed line which represents the ROC of a classifier that makes random predictions. Errors are too small to be shown.

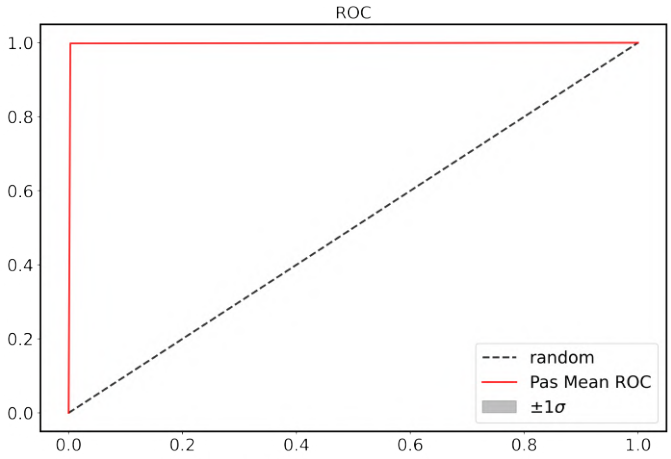


Figure 4.4: The ROC curve for the class of passive (Pas) galaxies (red solid line). The area under the curve (AUC) is above 0.99. The calculation of the mean ROC performed with the k-fold (k=10) cross-validation method. The error is the 1σ of the standard deviation of the k scores. For comparison, we plot the black dashed line which represents the ROC of a classifier that makes random predictions. Errors are too small to be shown.

4.2 Reason of success

In section 4.1 we estimated the performance of our new diagnostic on unseen data by adopting various metrics. All performance scores that were nearly perfect indicating that the diagnostic tool we defined can actually discriminate galaxies based on the three principal gas excitation mechanisms very well.

Firstly, the success of the diagnostic can be attributed to the selection of key features. The three EW values have distinct distributions for each one of the three principal galaxy classes and are characteristic of the activity of each class. As seen from the feature distributions (Figure 2.5) the EW of Hα for the passive galaxies are concentrated around 0. The same trend for the passive galaxies is true for the other two EW we have included here. This can be explained by the nature of passive galaxies since they have already used most of their gas leaving them depleted from gas and dust. Furthermore, the EW of the AGN galaxies in the forbidden line of [OIII] λ5007 have more negative values (negative means emission) than the other classes.

Furthermore, we have selected the D4000 continuum break index as an indicator of the presence of hot evolved stellar populations in a galaxy. In Figure 2.5 it is clear that the distribution of the D4000 for the SF galaxies have a very distinct distribution when compared to selected sample of passive galaxies. We have to remember that the sample of passive galaxies we defined here contains not only red, retired galaxies with no activity at all, but also objects with weak emission lines. In fact, this reveals that the D4000 index can be used as a good indicator that helps to separate excitation from young stars from the one produced as a result of old stellar populations. That is possible as the D4000 break is affected by two main things. The first is that as the galaxy ages the increasing lack of blue stars makes its continuum, which is produced by the superposition of many black bodies, stepper when compared to what it was when the galaxy was younger. In addition, the atmospheres of old stars are rich in metals which absorb more high energy photons creating a stepper break in the blue part of the spectrum. Thus we expect passive galaxies to show higher values of in the D4000 than the SF galaxies.

In section 3 we described how the Random Forest algorithm works. In particular, it was mentioned that the algorithm can provide us with the feature importance. In Figure 4.5 the feature importance plot that was produced during the training for the definition of the new diagnostic tool is presented. It is obvious that the EW value of Hα is ranked as the more important feature for the discrimination of galaxies in three principal activity classes. Also, we see that the D4000 is ranked as the second most important feature. The results from the feature importance supports our feature scheme choice because we can see that all features are of almost equal relevance. Also, we see that all features are almost of equal importance, which means that no feature is redundant.

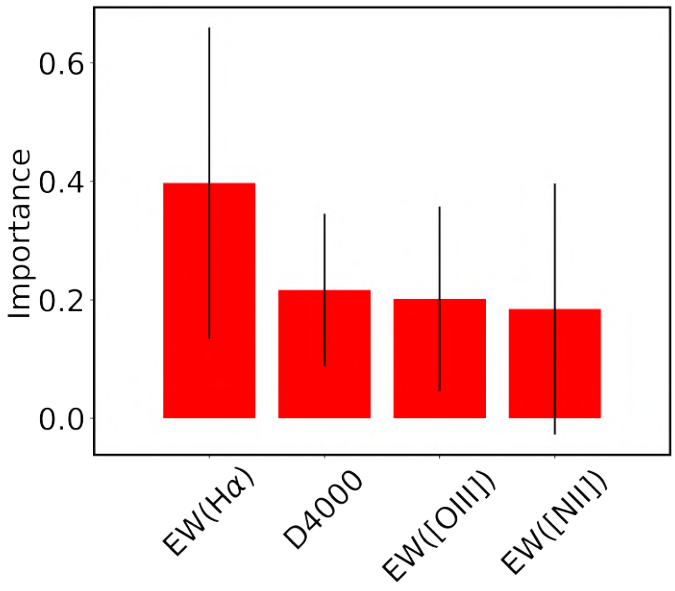


Figure 4.5: Plot of the feature importance for the calculated during the training of the algorithm. The error bars represent the standard deviation.

4.3 Classification of composite and LINER galaxies

In this work, in order to define our new diagnostic, we only considered three classes that are representative of the principal mechanisms of the gas excitation. However, there are galaxies where more than one principal gas excitation mechanisms can coexist. This is the example of the composite galaxies. In this situation a galaxy can have an active nucleus while star-formation processes are still present. Another possible scenario for a composite galaxy is that their gas excitation mechanism can

be from hot evolved stars while there is also an ongoing star-formation component. Additionally, the class of LINER galaxies is another mixed activity class of galaxies that their activity can be resolved into a combination of the three principal activity mechanisms. The origin of their activity is still a topic of controversy. As proposed by Ho, Filippenko, and Sargent, 1997 LINER galaxies can be separated into two categories type 1 and type 2. The type 1 LINERs are galaxies that their emission spectrum can be explained by a low-luminosity AGN (LLAGN), as many galaxies of this type appear to have broadened lines indicating the presence of an accretion process. On the other hand, there is mounting evidence that type 2 LINERs are the result of excitation from hot evolved stars (Binette et al., 1994;Stasińska et al., 2008;Papaderos et al., 2013).

The approach we followed for the definition of our new activity diagnostic tool gave as excellent performance results for all principal activity classes. The training sample of the principal activity classes have well-defined distributions in this 4-dimensional feature space with extremely good separation, while at the same time they have same overlap. This overlap is crucial for the mapping the full extent of the distributions of the different classes in the 4-dimensional feature space and to identify outliers in each of the three classes.

This means that the diagnostic can also classify successfully galaxies that belong to mixed activity classes based on their similarities to each one of the pure activity classes. Moving one step further, we can decompose mixed activity classes into their principal activity components. This decomposition ability of the classifier is a direct result of the chosen feature space since mixed classes will exist in the intermediate space between the principal classes.

This activity decomposition is achieved through the predicted probabilities of the Random Forest. When a the diagnostic is applied on a galaxy of a mixed class, each decision tree votes for the class that the object under question resembles more. As it is natural this object will share similarities with more that one principal class leading to lower maximum predicted probability for its first ranked class, which can be similar with another or two other classes.

During the processes of model evaluation, we observed that the classes of galaxies are already very well separated. This can be deduced not only by the great performance scores achieved by the application on the test data but also the predicted probabilities of the same data. This observation leads to the conclusion that we can not use this method for transforming the raw probabilities to absolute calibrated probabilities as there is not sufficient number of objects with intermediate estimated probabilities to correctly map the probability space. Thus, in order to avoid any biases that could be introduced by the probability calibration process, we use as predicted probabilities the raw probabilities as they are calculated directly from the algorithm.

Afterwards, we can use the output probabilities from the Random Forest to define the selection criteria for each class and also to identify the most likely classes. For this reason we use the probability of the most likely class (maximum probability max_p_i) and the difference between the probability of the first and the second ranked class for each object (Δp). By further analysing the predicted probabilities on the test sample for each of the three principal classes individually, we find that 90% of the population of the star-forming and passive galaxies have maximum predicted probabilities (max_p_i) above 90%. Because the precision and the recall for the two classes is ~ 1 this probability actually represents represents the 90% of the population of these two classes. Effectively, this means that the diagnostic has recovered every object of these two classes without including any objects from other classes. The slightly lower precision score for the class of AGN galaxies suggests a small amount of contamination. So, in order to find the probability limit that delineates the true AGN population we remove any SF and passive galaxy that was misclassified as AGN before estimating the minimum predicted probability for 90% of the true AGN population (the 90% fraction is adopted for consistency with the SF and passive galaxy classes). This estimation leads us to the conclusion that 90% of the true AGN population on the test set has maximum (max_p_i) predicted probability above 90%. A further verification of the validity of this probability limit, is that almost all misclassifications have max_p_i well bellow the 90%. For this reason we define as selection criterion for the three classes of max_$p_i \geq 90\%$. In other words, any galaxy that is predicted to belong to a class with a probability

higher than 90% is considered to belong in one of the classes of SF, AGN or passive. We describe these cases as being dominated purely by only one principal activity component.

However, as it was stated earlier, some galaxies do not host a unique gas excitation mechanism. In some cases there could be two competing ones. These galaxies will present characteristics in their observed features that will resemble more than one principal activity class and the application of the classifier on them will result in a low probability to belong to a given class (max_p_i) while its predicted probability to belong to each one of the other classes will be elevated compared to the galaxies that are dominated by only one principal excitation mechanism. In particular, we expect the majority of composite and LINER galaxies to have max_p_i < 90% while at the same time having comparable probabilities to belong on two of the three principal activity classes.

Taking it a step further, we can take into our consideration the predicted probabilities for each of the considered classes for an individual galaxy. By examining the two highest predicted probabilities we can broaden the number of predicted classes. More specifically, if a mixed activity galaxy (max_p_i < 90%) is predicted to have one of the two highest probabilities to be SF then the galaxy is considered as mixed starburst. Then, depending probability of the other competing class it can be either starburst-AGN or starburst-passive, if the competing probability is AGN or passive respectively. For example, a galaxy that has predicted probabilities to be a SF and AGN that are comparable while the probability of being passive is very small, then this galaxy is considered as starburst-AGN galaxy. Accordingly, we can define additional class if we consider the nature of LINERs. Again, this is a class of mixed activity, therefore we expect two of the predicted probabilities to have similar values. The emission of a LINER galaxy can be dominated by a low-luminosity AGN, hot evolved stars or both. Thus, the two highest class probabilities we expect to dominate are those probabilities to be AGN or passive. This means that a galaxy that has its two highest predicted probabilities to be AGN or passive the galaxy is characterised as passive-AGN.

Furthermore, by taking into account the principal activity class that holds the maximum predicted probability, we can refine our classification scheme further. Apart from the two principal class labels that appear in the name of each class, the order of appearance also matters. To be more specific, every mixed activity class label is characterised by two principal activity classes. The combination of these contributing classes describes the two dominant gas excitation mechanisms that are found in each mixed activity class. For this reason we define a new characterization scheme where, on the label of a mixed activity class we place in the first position the principal class with the highest predicted probability and the second identifier is the class with the second probability. For example, under this refined classification scheme the mixed activity class of starburst-AGN is different than AGN-starburst. Even though, both of them describe that the activity of these two galaxy classes are characterised mainly by star-formation and AGN process, the class of starburst-AGN describes a galaxy that the dominating source of excitation comes from star-formation while the class of AGN-starburst describes a galaxy that the dominating source of excitation is a result of an active nucleus. We note that the dominant source of ionization is determined from the similarity of an object with AGN, SF, and passive galaxies in the 4-dimensional space we consider here, and not on the dominant flux of ionizing photons based on SED analysis. Complete definitions and selection criteria about the classes of the more refined activity classification scheme can be found on tables 4.2 and 4.3 respectively.

Besides all the above mentioned cases, there could be galaxies that their predicted probabilities will be equally distributed between the three principal classes. In this case we have to establish a reliability selection criterion. This criterion will ensure that there could be one or two competing principal activity classes. The problem of an abject having comparable probabilities in all three classes is more prominent in the mixed activity objects. Thus, the sample we use to find a reliability threshold contains only composite and LINER galaxies. We implement our diagnostic on this sample to obtain their predicted probabilities to belong in one of the three principal classes. Afterwards, we calculate the difference between the maximum predicted and the second higher probability (Δp) as well as the difference between the second higher and the lowest predicted probability ($\Delta p'$) for each object. In Figure 4.6 we plot the Δp against $\Delta p'$. This is the probability difference between the first and the

Table 4.2: Definitions of the refined activity classes.

Class name	Definition
pure-starburst	The dominant photo-ionization source is young massive stars (star-forming).
pure-AGN	The dominant photo-ionization source is an active nucleus.
pure-passive	The dominant photo-ionization source is hot evolved stellar populations.
starburst-AGN	Two prevailing excitation sources, star-formation and an active nucleus. The dominant of the two is star-formation.
AGN-starburst	Two prevailing excitation sources, star-formation and an active nucleus. The dominant of the two is an active nucleus.
starburst-passive	Two prevailing excitation sources, star-formation and hot evolved stars. The dominant of the two is star-formation.
passive-starburst	Two prevailing excitation sources, star-formation and hot evolved stars. The dominant of the two is hot evolved stars.
AGN-passive	Two prevailing excitation sources, hot evolved stars and an active nucleus. The dominant of the two is an active nucleus.
passive-AGN	Two prevailing excitation sources, hot evolved stars and an active nucleus. The dominant of the two is hot evolved stars.
inconclusive	All three classes have similar probabilities

Table 4.3: New activity classes that include specific information about mixed classes. The max_p$_i$ is the probability of the highest ranking class for a galaxy assigned by the Random Forest classifier.

Class name	Criterion
pure-starburst	max_p$_i \geq 90\%$ to be SF
pure-AGN	max_p$_i \geq 90\%$to be AGN
pure-passive	max_p$_i \geq 90\%$ to be passive
starburst-AGN	max_p$_i < 90\%$ to be SF, while the second higher predicted probability is AGN.
AGN-starburst	max_p$_i < 90\%$ to be AGN, while the second higher predicted probability is SF.
starburst-passive	max_p$_i < 90\%$ to be SF, while the second higher predicted probability is passive.
passive-starburst	max_p$_i < 90\%$ to be passive, while the second higher predicted probability is SF.
AGN-passive	max_p$_i < 90\%$ to be AGN, while the second higher predicted probability is passive.
passive-AGN	max_p$_i < 90\%$ to be passive, while the second higher predicted probability is AGN.
inconclusive	Any galaxy satisfying the equation $\Delta p < -2 \cdot \Delta p' + 0.8$

second class plotted against the probability difference between the second and the third class. We note that the ranking of the classes in each object is different; in this analysis we are only interested in the probability difference as a metric of the discriminating power of the method and not the actual classes.

By plotting the Δp against $\Delta p'$ we can see that the bottom left corner of this plot is populated by objects with comparable Δp and $\Delta p'$ values. We characterise these objects as not reliably classified as these are having similar probabilities to belong in all three classes. Furthermore, we see that no object passes the line of $\Delta p = -2 \cdot \Delta p' + 1$. This is direct consequence of having three classes in total since these three predicted probabilities must sum to 1 for each object. This line sets the upper limit of an object in the Δp-$\Delta p'$ space. We consider this equation as the extreme reliability line, in other words, based on equation $\Delta p = -2 \cdot \Delta p' + 1$ defines the further distance from the origin which is the point

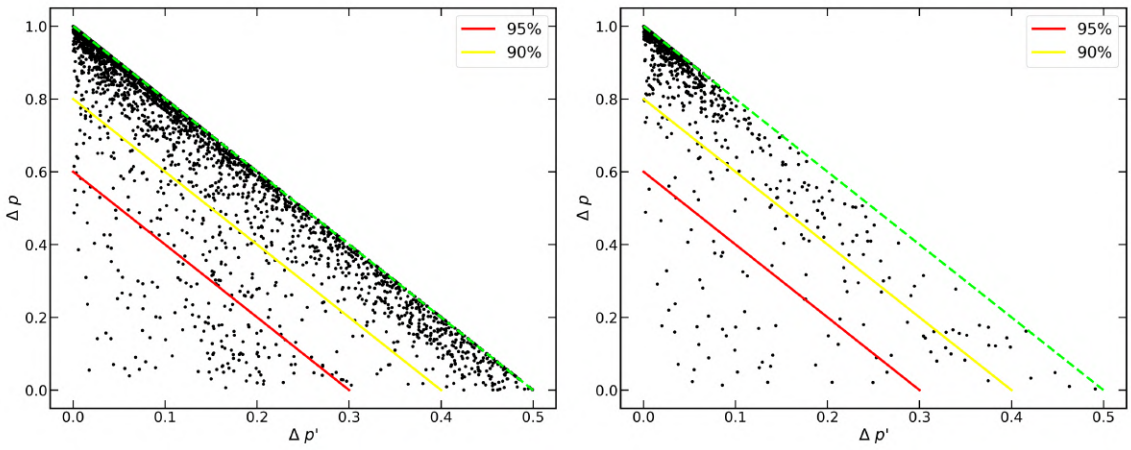


Figure 4.6: Plot of Δp against $\Delta p'$. In both plots, each black dot represents a galaxy of mixed class. The left plot is populated by the composite galaxies while the right plot is populated by the LINER galaxies. We see that the left bottom corner is populated by objects that have Δp=$\Delta p'$=0, which means that these objects are equally probable to belong to all classes resulting in unreliable classification. The green dashed line represents to the extreme reliability line (Δp = -2·$\Delta p'$ + 1), the region between the green line and the yellow solid line (Δp = -2·$\Delta p'$ + 0.8) contains 90% of the mixed class objects for both composites and LINERs and the region between the green and the solid red line (Δp = -2·$\Delta p'$ + 0.6) encloses 95% of the mixed class objects for both composites and LINERs.

of maximum mixing. Moving a line parallel to the extreme reliability line downwards towards the origin of the coordinate system we can find the combination of Δp-$\Delta p'$ probabilities that contain 90% of the objects. This effectively defines a reliability threshold that includes 90% of each population, in a similar way as the reliability thresholds defined for the pure classes (table 4.3). The equation has the same slope as the extreme reliability line by different intercept that depends on the scatter of the objects in the Δp, $\Delta p'$ plot. We find that the equation that satisfies the above criteria is the Δp = -2·$\Delta p'$ + 0.8. Thus, we mark every object that satisfies the relation of Δp < -2·$\Delta p'$ + 0.8 as having inconclusive classification. Our chosen reliability criterion is represented by the yellow solid line.

Taking into consideration this analysis and the results from the performance evaluation of our diagnostic, we use our diagnostic for decomposing the classes of composite and LINER galaxies into the principal components of the gas excitation mechanism.

This way, on a sample of composite and LINER galaxies. We acquire the sample of composite and LINER galaxies using the same catalog of galaxies, diagnostic tool, and applying S/N criteria as described in section 2 which was used to select the other two activity classes (SF and AGN galaxies). Then, we are going to classify these objects following two approaches. In the first approach, we apply the new diagnostic tool on the selected sample of composite and LINER galaxies in order to classify them in one of the three principal classes. This will allow us to classify these galaxies based on their similarity to one of the three principal classes. Based on the previous discussion a galaxy that is more similar to star-forming galaxies (i.e., the EW of the diagnostic lines falls closer to the locus of the SF galaxies) will be classified as a SF composite. Accordingly, we characterize the rest of the mixed activity galaxies as AGN-composite, passive-composite, SF-LINER, AGN-LINER and passive-LINER, with the first component of the name stating one of the three principal activity classes that resembles the most and the other its spectroscopic classification. In the second approach, we will classify the mixed activity galaxies based on the refined classification scheme that is thoroughly described on tables 4.3 and 4.2. Our goal with this two step approach is to analyse how the first, crude activity component decomposition becomes refined into more specific classes that describe better the activity characteristics of a galaxy.

After the application of our new diagnostic tool on the sample of composite and LINER galaxies

we can calculate the fractions of composite and LINER galaxies that have been predicted as being dominated by star-formation, AGN and hot evolved stellar emission. For the composite galaxies it is found that about 1/2 of them are predicted as being dominated by star-formation, 1/12 as being dominated by an active nucleus and 1/4 as being dominated by hot evolved star (all these fractions refer to the total number of objects of our sample of composite galaxies). For the case of LINER galaxies we find that about 1/8 are predicted as being dominated by AGN activity and 4/5 by hot-evolved stars. LINER galaxies predicted as star-forming dominated are almost non existent (all these fractions refer to the total number of objects of our sample of LINER galaxies). In the tables 4.4 and 4.5 we summarize accurately the latter results for the composite and LINER galaxies respectively.

Table 4.4: Class predictions after the application of the new diagnostic on the sample of composite galaxies. The first column represents the predicted class as given by the Random Forest based on the most probable class. The second and the third columns are the the percentage and the number of the objects to the total population of the spectroscopically classified composites respectively. Any object that satisfies $\Delta p < -2 \cdot \Delta p' + 0.8$ is labeled as inconclusive.

RF predicted class	Percentage (%)	Galaxies
SF-composite	55.4	1578
AGN-composite	8.4	238
Passive-composite	26.4	752
Inconclusive	9.8	280
Total	100.0	2848

Table 4.5: Class predictions after the application of the new diagnostic on the sample of LINER galaxies. The first column represents the predicted class as given by the Random Forest. The second and the third are the the percentage and the number of the objects respectively to the total population of the LINERs. Any object that satisfies $\Delta p < -2 \cdot \Delta p' + 0.8$ is labeled as inconclusive.

RF predicted class	Percentage (%)	Galaxies
SF-LINER	0.7	8
AGN-LINER	12.9	152
Passive-LINER	77.2	910
Inconclusive	9.2	108
Total	100.0	1178

We can analyse further these results by exploring the locus of these different subclasses on the standard emission line ratio (BPT diagnostic) diagrams. In that plot [OIII] λ 5007Å/Hβ) against [NII] λ 6584Å/Hα the composite galaxies occupy the area that is between the two lines of Kauffmann et al., 2003 and Kewley et al., 2001. This way, the results can be inspected from a different perspective as we transfer them to a 2-dimensional projection in a different feature space form the one used for the classification. In Figures 4.7 and 4.8 we present the projections of the predictions made by the new classifier for the characterization of the principal activity mechanism on a [OIII] λ 5007Å/Hβ) against [NII] λ 6584Å/Hα for the composite and LINER galaxies respectively. Starting from the composite galaxies it can be observed that the classes assigned by the new diagnostic tool have clouds that are clearly separated as the center of each distribution is distinguishably different from the other. On that diagram, the location of the composite galaxies which we find that are dominated by star-formation processes are just above the star-forming cloud and tangential to the Kauffmann et al., 2003 line. correspondingly, the AGN-composite predicted galaxies are found in the upper part of the composite population, close to the theoretical line of extreme starburst line defined by Kewley et al., 2001. For the case of the passive-composite galaxies we see that the distribution of these mixed activity class is more wide and elongated with a positive slope. Their location is close to the area of objects with strong low-ionization emission-lines (LINERs). It is particularly interesting that they follow the trend

described in Byler et al., 2019 for galaxies with contribution from an ionizing component of older hot stellar populations.

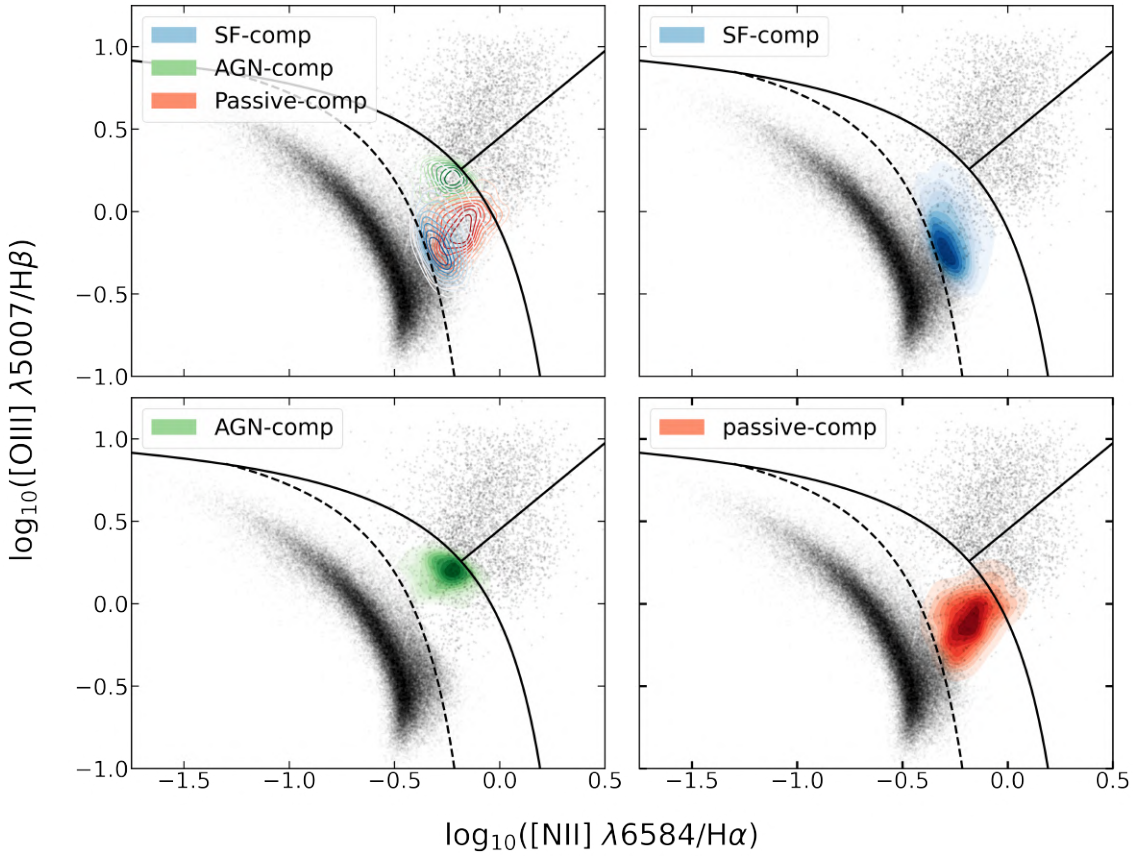
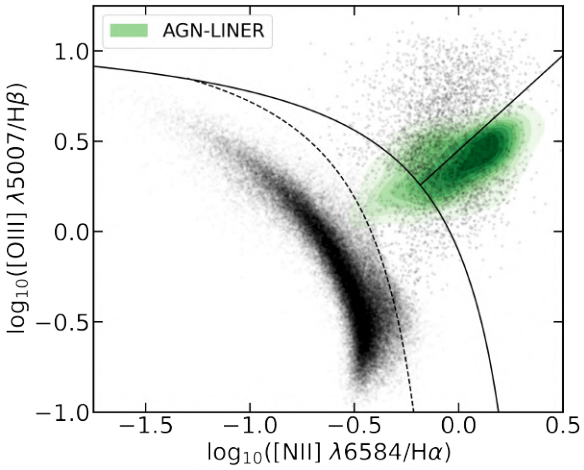
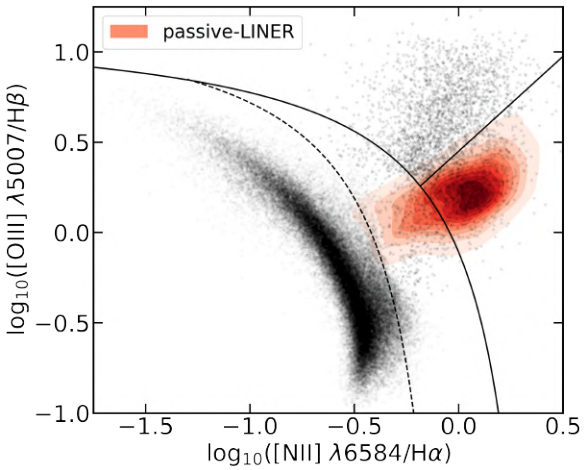


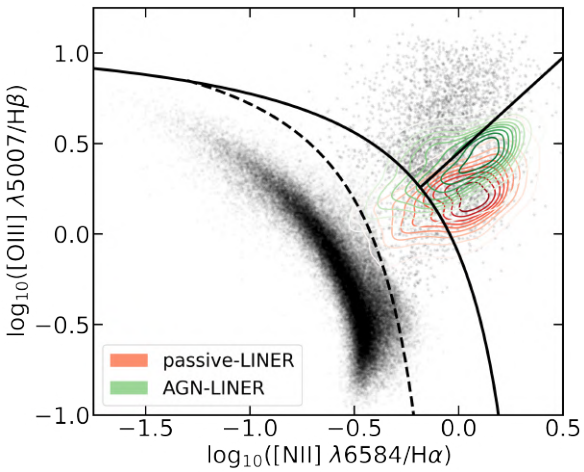
Figure 4.7: Four BPT plots of $\log_{10}$([OIII] λ 5007/Hβ) against $\log_{10}$([NII] λ 6584/Hα. Top left: Gaussian kernel density contours of the spectroscopically selected sample of composite galaxies that have been predicted to belong to one of the principal activity classes: Blue represents composites predicted as SF, green as AGN and red as passive. This plot shows the overlap of some SF and passive predicted composite is visible. The rest of the plots (top right, bottom left and bottom right) shows the density and the span of each of predicted class for the sample composites. The black dashed line is the Kauffmann et al., 2003. The black solid curved line is the Kewley et al., 2001 while the straight black line is the Schawinski et al., 2007 separating LINERs from AGN. The black dotes are the training sample shown for demonstration purposes.



(a)



(b)



(c)

Figure 4.8: Three BPT plots of $\log_{10}$([OIII] λ5007/Hβ) against $\log_{10}$([NII] λ6584/Hα. The first two plots, (a) and (b), shows the density and the span of each of predicted class for the sample of spectroscopically selected LINERs. (c): Gaussian kernel density contours for sample of LINER galaxies that have been predicted to belong to one of the principal activity classes: green represents LINERs that have been predicted as AGN and red as passive. This plot shows that there is some overlap between the LINERs predicted as AGN and as passive. The black dashed line is the Kauffmann et al., 2003. The black solid curved line is the Kewley et al., 2001 while the straight black line is the Schawinski et al., 2007. The black dots represent the training sample shown for demonstration purposes.

Implementing the second activity decomposition approach for the subsample of mixed activity galaxies allows us to perform a more refined classification. As we described earlier, we are going to classify the set of mixed classification galaxies into ten classes described in detail in tables 4.2 and 4.3. In Figures 4.9 and 4.10 we present two histograms that present the percentages of the composite galaxies that belong to each subclass. In the first histogram (Figure 4.9) we show the predictions based on the likelihood of similarity of a composite galaxy to one of the three principal activity class. In the next histogram (Figure 4.10), we show the classification based on the more refined activity classification scheme. Comparing these two histograms we deduce that the majority of composite galaxies are actually the result of a combination of two principal activity classes. In Figure 4.11 we project these composite galaxies onto the standard BPT plot, in order to see the location of each one of the refined activity classes.

We can repeat the same procedure on the subsample of LINER galaxies. In Figures 4.12 and 4.13 we present the results of the classification obtained by discriminating them based on their similarity to the three principal activity classes and with the refined classification scheme that also considers the mixing of the different gas excitation mechanisms present in the host galaxy. In Figure 4.14 we plot the subsample of LINER galaxies on a standard BPT plot. It is known that the [SII] doublet and the [OI] are good probes of low-ionization sources. For this reason, in Figure 4.15 we plot the subsample of LINERs on a [OIII]/Hβ against [SII]/Hα and on a [OIII]/Hβ against [OI]/Hα plots. The classification labels we use were assigned by the refined activity model.

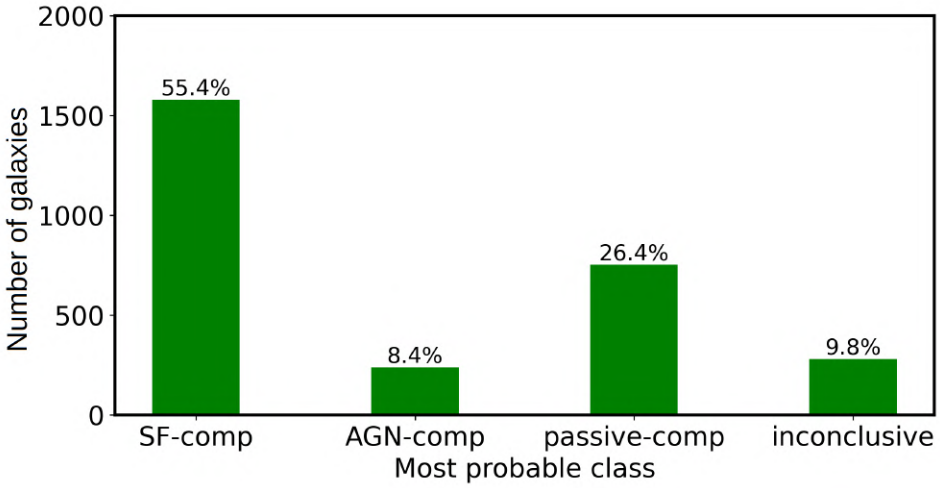


Figure 4.9: Histogram for the most probable, highest similarity class with one of the three principal activity classes for the subsample of composite galaxies. The objects marked as inconclusive, are objects that satisfy the equation $\Delta p < -2 \cdot \Delta p' + 0.8$ and thus the result of their classification is considered as unreliable.

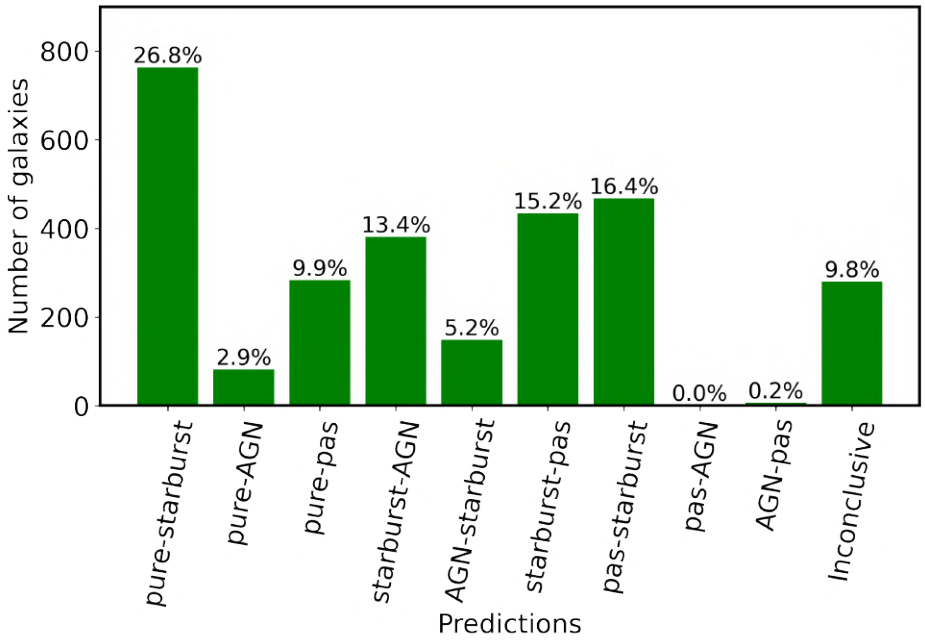


Figure 4.10: Histogram of the same objects that appeared on Figure 4.9 but in a more refined classification. This is achieved by considering not only the predicted class (class with maximum probability) but also the second higher predicted probability, providing us with more information about the actual origin of activity of a particular galaxy. The objects marked as inconclusive, are objects that satisfy the equation $\Delta p < -2 \cdot \Delta p' + 0.8$ and thus the result of their classification is considered as unreliable.

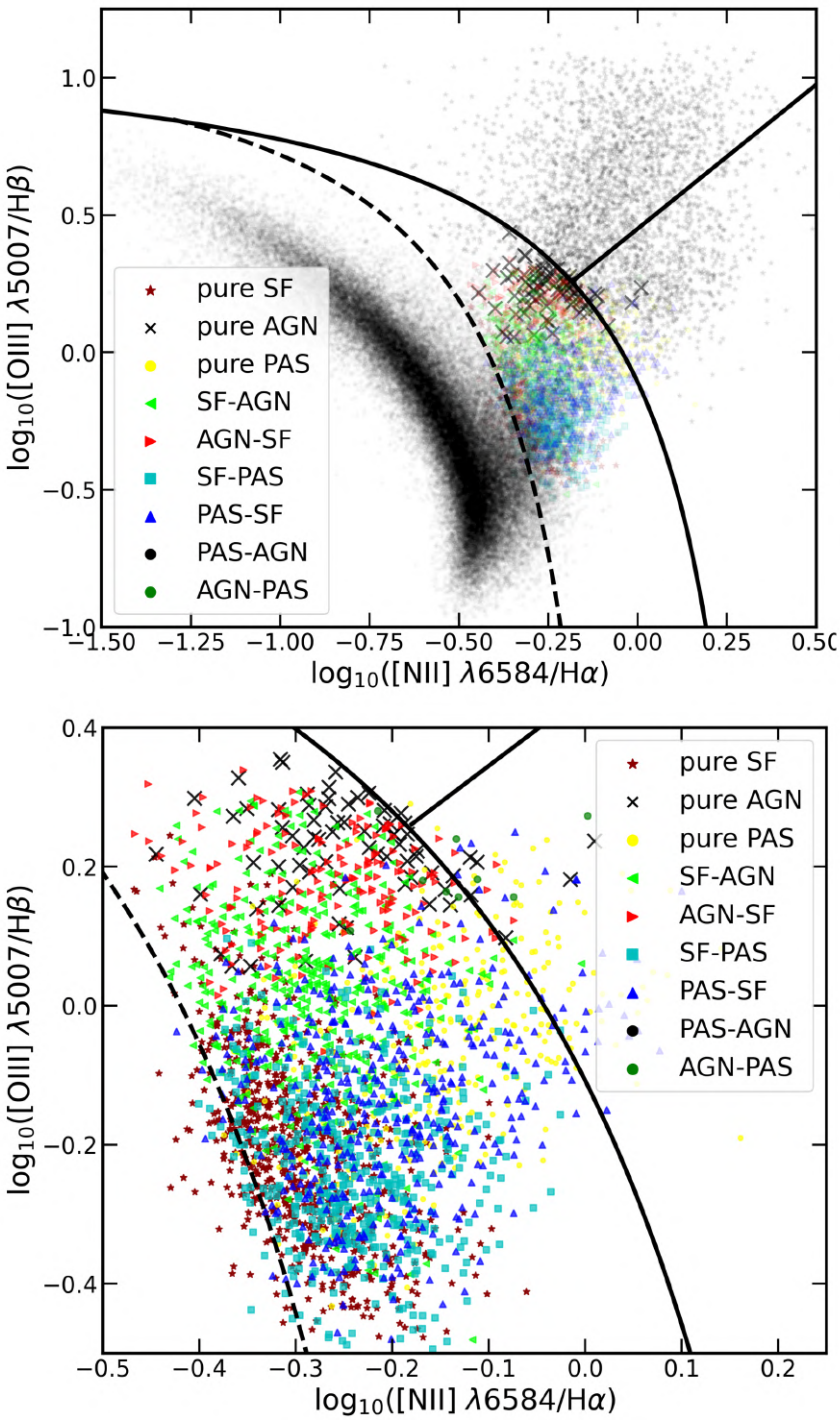


Figure 4.11: Two standard BPT plots, [OIII]/Hβ against [NII]/Hα. The top plot shows the location of the refined activity predictions for the spectroscopically selected subsample of composite galaxies. On the bottom plot we see the same objects on a close up view of the overlap region that better highlights our results. The black dashed line is the Kauffmann et al., 2003 line separating star forming from composite galaxies. The black solid curved line is the Kewley et al., 2001 which is the theoretical extreme starburst line, while the straight black line is the Schawinski et al., 2007 separating LINERs from AGN. The black dots on the top plot show the training sample of the principal classes shown for demonstration purposes.

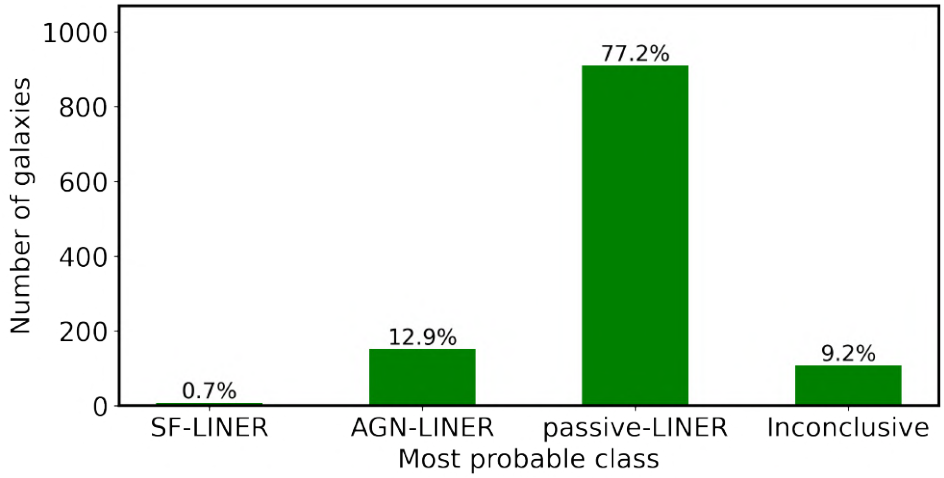


Figure 4.12: Histogram for the most probable, highest similarity class with one of the three principal activity classes for the subsample of LINER galaxies. The objects marked as inconclusive, are objects that satisfy the equation $\Delta p < -2 \cdot \Delta p' + 0.8$ and thus the result of their classification is considered as unreliable.

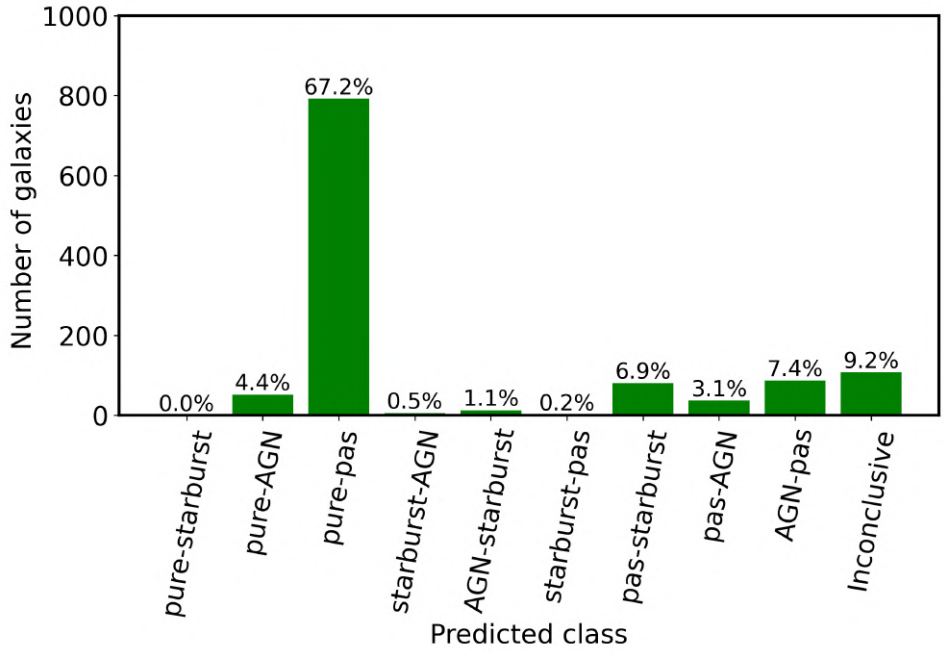


Figure 4.13: Histogram of the same objects that appeared on Figure 4.12 but in a more refined classification. This is achieved by considering not only the predicted class (class with maximum probability) but also the second higher predicted probability, providing us with more information about the actual origin of activity of a particular galaxy. The objects marked as inconclusive, are objects that satisfy the equation $\Delta p < -2 \cdot \Delta p' + 0.8$ and thus the result of their classification is considered as unreliable.

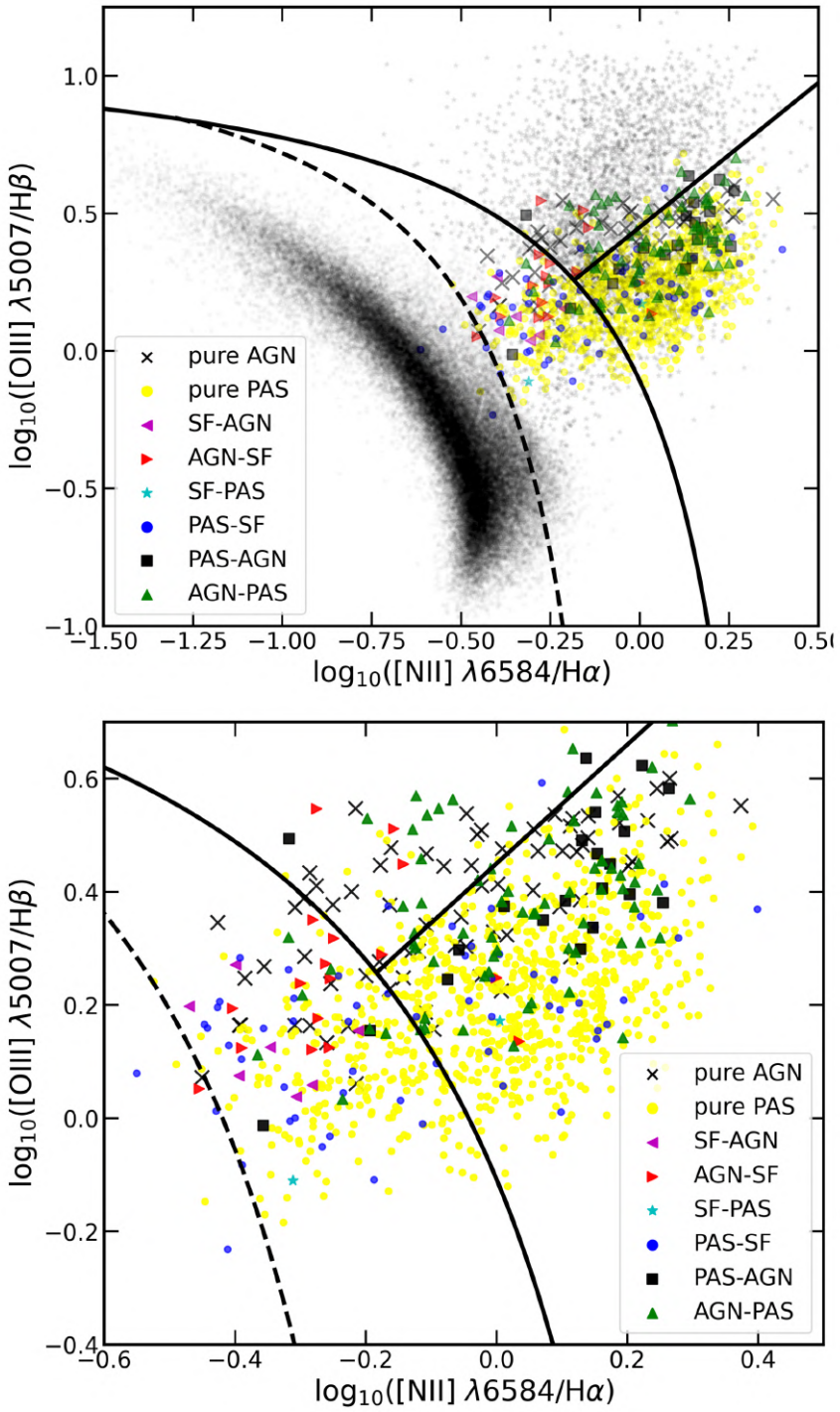


Figure 4.14: Two standard BPT plots, [OIII]/Hβ against [NII]/Hα. The top plot we can see the location of the refined activity predictions for the spectroscopically selected subsample of LINER galaxies. On the bottom plot we see the same objects on the same plot as on top but in a close up view that better highlights our results. The black dashed line is the Kauffmann et al., 2003 line separating star forming from composite galaxies. The black solid curved line is the Kewley et al., 2001 of extreme starburst while the straight black line is the Schawinski et al., 2007 separating LINERs from AGN. The black dots on the top plot show the training sample of the principal classes shown for demonstration purposes.

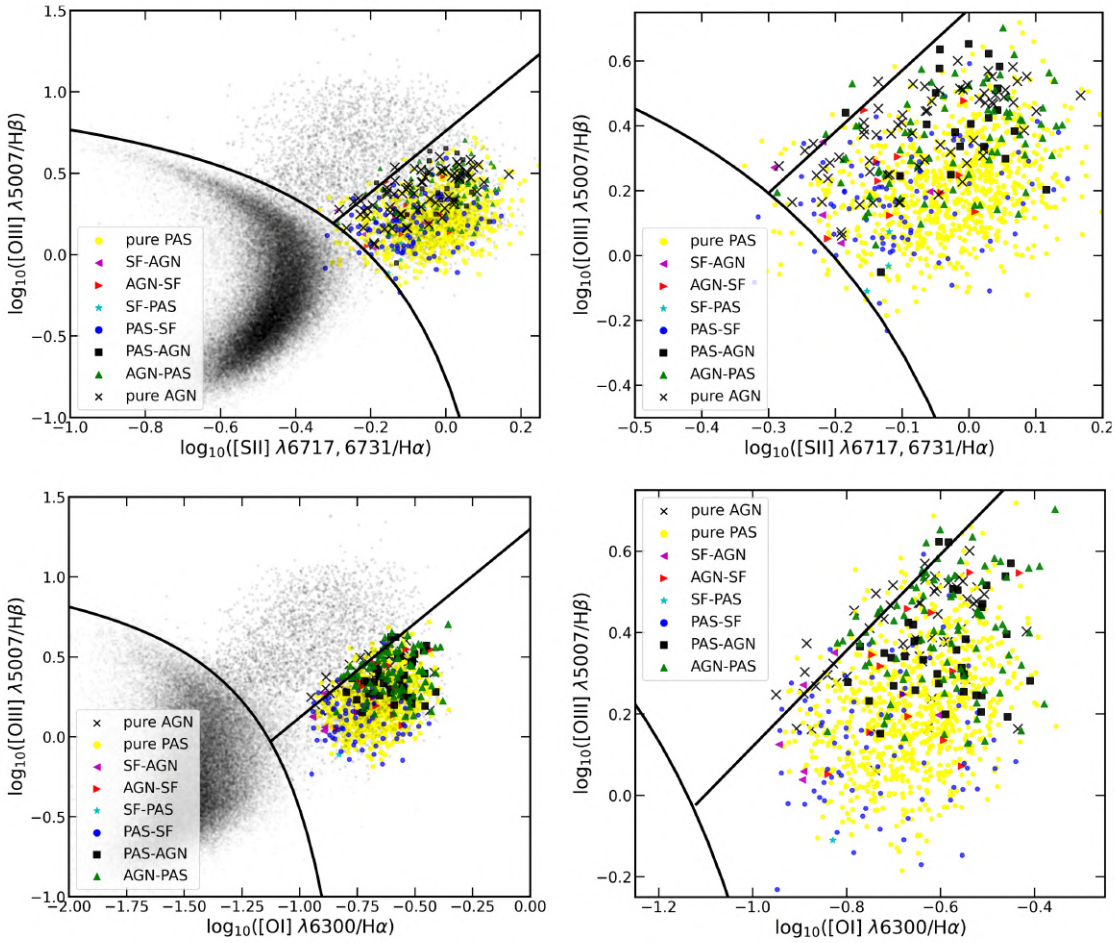


Figure 4.15: Top row: two plots of [OIII]/Hβ against [SII]/Hα for the spectroscopically selected subsample of LINER galaxies. The left plot includes the training sample (black dots) for better visualisation of the position of each class. The right plot is the same plot focusing on the region of LINERs. Bottom row: plots of [OIII]/Hβ against [OI]/Hα for the same subsample of LINER galaxies. The left plot includes the training sample (black dots) for better visualisation of the position of each class. The right plot is the same plot offering a better resolution of the position of each refined activity class. In all cases, the classification labels are based on the refined activity classes. In all four plots the black solid curved line is the Kewley et al., 2001 separating star-forming from AGN and LINERs while the black solid straight line is the Schawinski et al., 2007 separating AGN from LINERs.

5. Discussion

5.1 Dominant photo-ionization mechanism of the host galaxy

In this work we considered only three galaxy classes: SF, AGN and passive. These classes represent the main types of activity that can contribute to the emission spectrum of a galaxy. This new diagnostic was defined based on the concept that the Equivalent Widths of the lines of Hα, [NII] λ6584Å, [OIII] λ5007Å combined with the D4000 index are sufficient to discriminate galaxies according to these three main types of activity as well as to separate excitation that is a result of old stellar populations form active star-formation. As we showed in the previous sections this is actually possible. The high scores over all performance metrics demonstrate the success of this effort.

Galaxies that are in a specific stage of their evolution (SF, AGN or passive) seem that they only host a specific type of activity, making their optical observed spectrum unique. However, as galaxies evolve their activity can not be attributed uniquely to one type of activity, thus their observed spectrum is usually a combination of the three principal activity classes. A good example of intermediate stages of a galaxy evolution that is difficult to identify the dominant component that drives their activity are the composite galaxies. The observed optical spectrum of these galaxies can be a combination of SF and AGN or SF and populations of hot evolved stars. The results obtained throughout the analysis support the idea that hot evolved stars can have significant contribution on the observed spectrum of a composite galaxy. Another interesting fact is that there is overlap between the SF-composite and the passive-composite galaxies at the bottom right area of the BPT diagram, just above the Kauffmann et al., 2003 line. This happens because the hot evolved stars sometimes can mimic the activity of an active galaxy (Stasińska et al., 2008) and the optical emission-lines are not sufficient for such discrimination. This potentially means that in the same area of the BPT diagram these two sub-populations of composite galaxies could coexist. The discriminating power offered by our diagnostic and which is crucial for this discrimination can be attributed to the D4000 break. As we can tell from the feature distributions Figure 2.5, the SF galaxies and the passive galaxies, unsurprisingly, have clearly separated distributions concerning the D4000 feature.

Considering the other complex activity class that is often considered as separate, the LINERs, we can see that we can also separate them in two subclasses based on their origin of activity. From Figure 4.8 it is clear that the LINER galaxies are separated into two distinct sub-populations. The population that has more similarities with the AGN galaxies lies close to the separation line of Seyfert and LINER galaxies defined by Schawinski et al., 2007. A remarkable fact that further supports our results is that, although there has been defined a clear separation line between AGN and LINER galaxies on the BPT diagram, Ho, Filippenko, and Sargent, 2003 emphasize that the separation line does not have an absolute physical significance. The distribution of AGN galaxies spans in a wider range than generally are considered to be. An other interesting characteristic of the distribution LINER galaxies classified as AGN is that its shape is elongated and has a slope that is parallel to the Seyfert-LINER separation line. The rest of the LINERs are predicted as passive. Their distribution is located beneath the cloud of the AGN predicted LINERs.

Another interesting fact for the composites that have been predicted as passive are the diagrams of

the low-ionization line ratios of [SII]/Hα and [OI] λ 6300Å/Hα. This is because hot evolved stellar populations are typically found in older galaxies with low gas reservoirs. In Figure 5.1 we see that indeed the composite galaxies that have been predicted as dominated by characteristics of passive galaxies are found bellow the extrapolation of the Schawinski et al., 2007 line while the composite galaxies that have been predicted as AGN are located above it. Moreover, the separation between the passive-composite and the AGN-composite is clear.

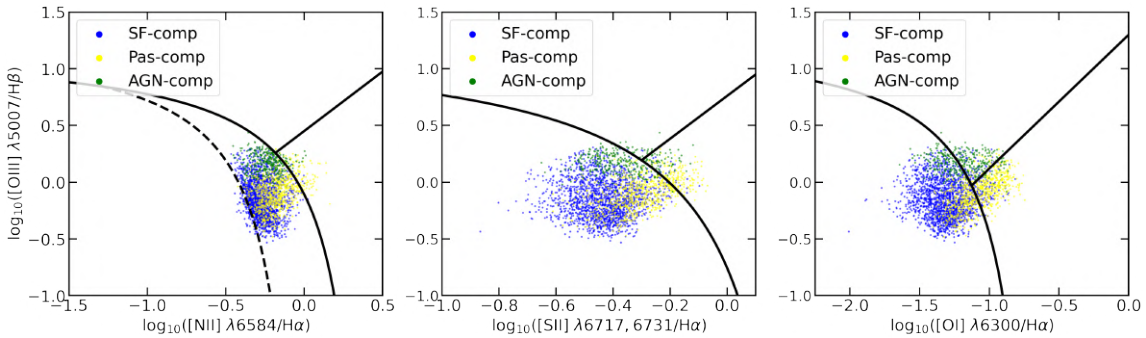


Figure 5.1: Three plots form left to right, [OIII]/Hβ against [NII]/Hα [OIII]/Hβ against [SII]/Hα and [OIII]/Hβ against [OI]/Hα for the spectroscopically selected sample of composite galaxies.

In Figure 4.7, we saw that some composite galaxies that were predicted as passive-composites are located just above the Kauffmann et al., 2003 line, when projected on the standard BPT diagram. This result may seem unnatural at first but it is actually in accordance with the recent work of Byler et al., 2019. In that work, the authors used photoionization models to show that the contribution of increasing age of hot evolved stellar populations displaces galaxies towards the bottom area in the standard BPT plot. This trend becomes more clear in Figure 5.2 which contains the distribution of the activity class of passive-composites alongside the track of the points from Byler et al., 2019. There we observe that the designated area proposed by the Byler et al., 2019 is mainly populated by galaxies that have significant contribution by old stellar populations (i.e, passive-starburst and starburst-passive galaxies). We note that the AGN-composite galaxies lie above the locus of the passive-composite galaxies shown here, while the SF-composite galaxies are located closer to the dashed line indicating the empirical line of Kauffmann et al., 2003 delineating the SF galaxies (Figure 4.7).

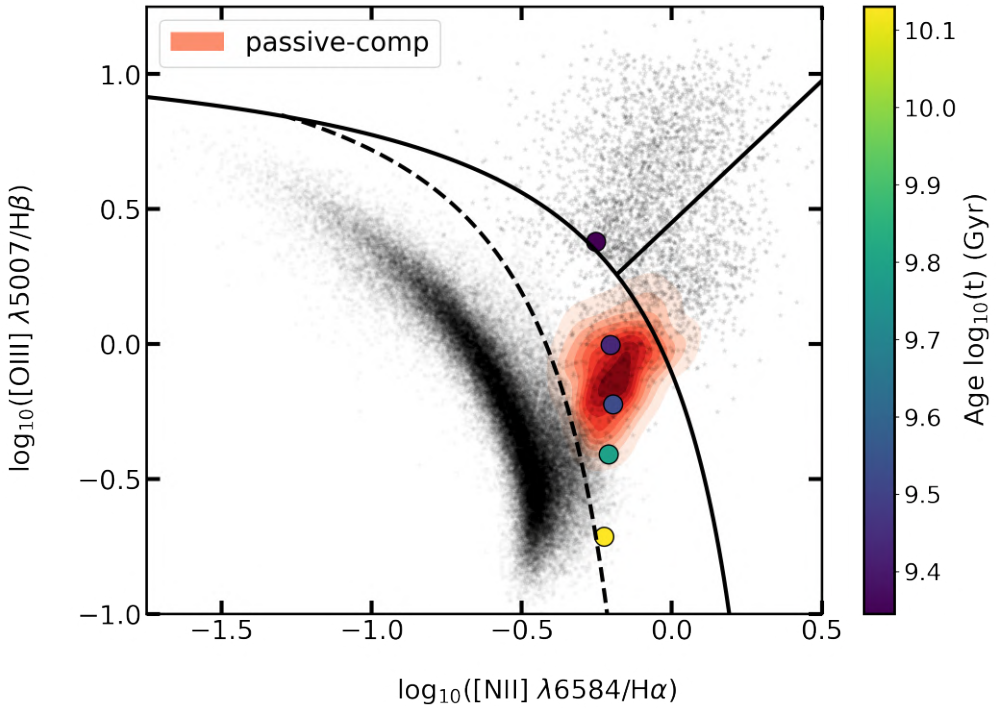


Figure 5.2: Standard BPT plot, [OIII]/Hβ against [NII]/Hα. In that plot we can see the location of the spectroscopically selected subsample of composite galaxies that have been classified by the new diagnostic as passive-composite galaxies (composites that the source of ionization comes from hot evolved stars). The circles are points derived from the work of Byler et al., 2019. We plot the circles to indicate the location of a galaxy with aging stellar populations form 2 to 14 Gyr. The circles are color-coded to represent age. The black dashed line is the Kauffmann et al., 2003. The black solid curved line is the Kewley et al., 2001 while the strait black line is the Schawinski et al., 2007 line separating AGN from LINERs. The black dots on the top plot are the training sample of the principal classes shown for demonstration purposes.

In conclusion, based on the facts discussed above this new diagnostic tool can be used not only for the classification of galaxies but also the characterization of the underlying activity of mixed galaxies. However, it should be noted that the probabilities obtained from our analysis do not correspond to the actual fractions of the contribution of each one of the corresponding principal activity mechanism to the observed spectrum instead it is the likelihood of similarity.

5.2 Pure class selection thresholds

In order to define the selection thresholds for the pure classes, we calculated the minimum predicted probability above of which is the 90% for each of the population of the three principal classes on the test set. This way we optimize the completeness of the pure class samples and minimize the possible missclassifications. In fact our analysis showed that a probability selection threshold of 90% for pure classes is the best as a higher one would result into a poorer recall score of the pure AGN class. This happens as the AGN generally have lower predicted probabilities when the diagnostic is applied on the test set. This is a result of some mixing between the classes of pure SF and pure AGN. This very small mixing is enough to lower the predicted probabilities of the pure AGN population but not significant to lower their performance scores (the maximum predicted probability remains AGN)

Even though the probability selection threshold of 90% was chosen based on the 90% population, other selection criteria were also considered. For example, another probability threshold was considered based on the 95% of the population of the pure classes. This is a more strict criterion. As shown in the Figure 5.3, the advantages of adopting a more strict selection threshold is that the

scatter is reduced and the composite area of the BPT contains fewer composite galaxies that are classified as pure starburst galaxies. The disadvantage, though, is that the recall of the pure AGN drops as the probability selection threshold increases, reducing the completeness of the AGN sample. Regarding the composite galaxies classified as pure star-forming, even though they are above the Kauffmann et al., 2003 line, we want to stress that although these galaxies have been classified as being mainly dominated by star-formation process their predicted probabilities are very close to the lower limit of the selection threshold (i.e., 90% for the adopted scheme or 95% percent for the more strict scheme). This is expected for two reasons. The first one is that the plot presented in Figure 5.3 is a projection from the 4-dimensional EW and D4000 feature space to the 2-dimensional feature space of the emission line ratios and as a result some scatter is expected. The second is that the transition from pure starburst to a mixed activity galaxy is a continuous process and happens gradually. The latter is also supported by the decreasing probabilities as we move inside the center of the composite area on a BPT diagram.

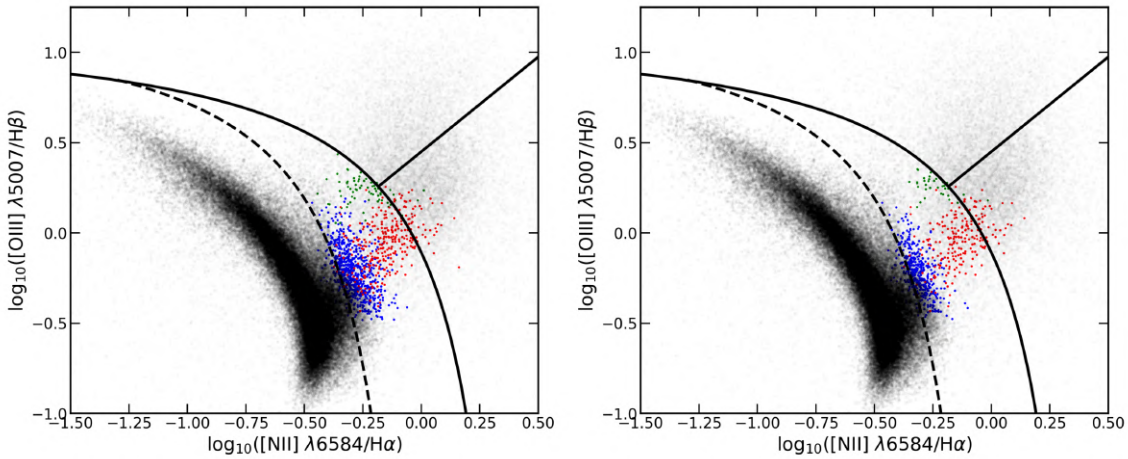


Figure 5.3: Comparison of the different selection thresholds for the pure classes. The criterion for a galaxy to be predicted to belong in a pure class is based on the maximum predicted probability given by the classifier. On the left, a BPT diagram [OIII] 5007Å/Hβ against the [NII] 6584Å/Hα showcasing the location where composite galaxies have been predicted to belong uniquely (pure classes) in one of the three principal activity classes with a maximum predicted probability threshold of 90%. On the right we provide the same plot but the selection threshold for pure class has been set to 95%. Red points represent the composite galaxies that have been classified as SF, green points as AGN and red as passive. The black points are the training sample of galaxies shown for demonstration purposes. We see that in the first case (90% criterion) the locus of galaxies classified as pure star-forming extends well into the area of composite galaxies, whereas in the case of the more strict scenario stays more close to the separating line between star-forming and composite galaxies.

5.3 Inconclusive classifications

In section 4.3 we mentioned that not all galaxies will receive a conclusive classification. This issue arises only in the mixed activity classes, galaxies that are spectroscopically identified as composites and LINERs. This happens because in our chosen 4-dimensional feature space the pure classes are very well separated and as a result galaxies with a single excitation mechanism will be classified in one of the three principal activity classes with high confidence. These are the galaxies that are located along the upper right ridge of the Δp-$\Delta p'$ diagram (Figure 4.6). Mixed class objects will be below this ridge, and we find that objects that have probabilities in the upper 90% percentile of the population which as discussed earlier turns out to be a probability of 90% (for all classes) fulfill the criterion of $\Delta p < -2 \cdot \Delta p' + 0.8$. These are the mixed activity objects, objects below this line are considered as having inconclusive classifications since their probability to belong into any of the considered classes

are very similar. In Figure 5.4 we present the 3-dimensional projection of the 4-dimensional feature space. In this plot we can see that inconclusive classifications are located in an confined area that connects all three principal activity classes. Every inconclusive classification object has originated form either a composite or LINER galaxy. In addition, we have to remind that the training of the Random Forest happens by selecting random samples for the training subsample which, unavoidably, leads to a small degree of uncertainty in the predicted probabilities.

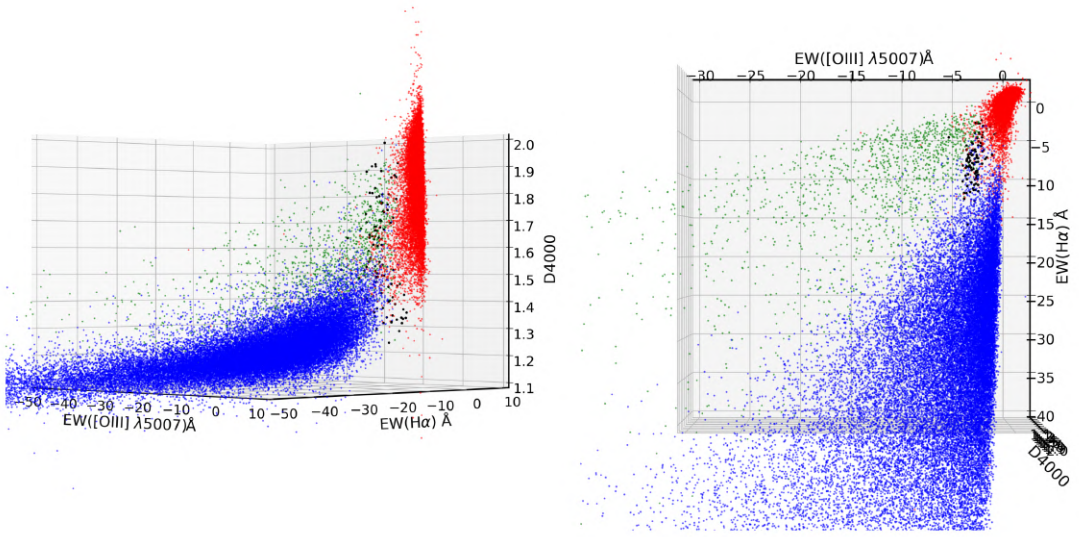


Figure 5.4: Two plots of the 3-dimensional projections of the 4-dimensional feature space used for the training of the new diagnostic. The blue dots are SF galaxies, the green dots are the AGN galaxies and the red dots are the passive galaxies. The black dots are galaxies that have been characterized as having unreliable classification. All black dots satisfy the criterion of $\Delta p < -2 \cdot \Delta p' + 0.8$, and they are located in between the reliable classes.

6. Conclusions

In this work we studied a classification problem that is based on the three principal activity mechanisms of the excitation (i.e., star-formation, AGN activity and photoionization by old stellar populations). We also studied the mixed activity classes of the composite and LINER galaxies in an attempt to characterize them based on the most likely source of their activity based on the three activity classes of SF, AGN and passive. We summarize our results and conclusions derived by this work bellow.

1. It is possible to define a simple diagnostic tool based on a machine-learning methods that uses only four spectral features and is capable of discriminating with high accuracy (0.983 ± 0.004) between the three principal activity classes, star-formation, active nucleus and excitation from old stellar populations. To our knowledge, this is the first time that a diagnostic tool manages to incorporate active and passive galaxies under one unified scheme while at the same time offering high reliability and completeness.
2. Galaxies can be separated based on the three principal activity mechanisms of gas excitation of star-formation, active nucleus and emission from hot evolved stars utilizing only the Equivalent Widths of the lines of Hα, [OIII] λ5007Å, [NII] λ6584Å and the of D4000 index.
3. We applied our new diagnostic tool successfully on the mixed activity classes of composite and LINER galaxies in order to identify the combination of the principal activity classes that are responsible for the gas excitation.
4. The application of our new diagnostic on the sample of composite galaxies resulted in partial overlap between the predicted SF-dominated composite and passive-dominated composite galaxies when projected on the 2-dimensional BPT diagram. This is a result of the degeneracy in the optical spectra (emission lines) between the excitation from star-formation and post-AGB stars. However, we find that it is possible to break this degeneracy with the inclusion of D4000 break as a discriminating feature.
5. We also find that we can discriminate LINERs into objects that are dominated by AGN activity and those with emission lines arising by photoionization by old stellar populations (post AGB-stars).
6. The probabilities that are provided by our diagnostic can be used as an indicator for the characterization of the principal activity mechanism of the host galaxy. However, these probabilities can be treated only as a likelihood of similarity and should not be used as a fraction of the actual contribution of each one of the principal mechanism to the observed spectrum.

Bibliography

[1] R. J. Assef et al. “Mid-infrared Selection of Active Galactic Nuclei with the Wide-field Infrared Survey Explorer. II. Properties of WISE-selected Active Galactic Nuclei in the NDWFS Boötes Field”. In: 772.1, 26 (July 2013), p. 26. DOI: 10.1088/0004-637X/772/1/26. arXiv: 1209.6055 [astro-ph.CO].

[2] J. A. Baldwin, M. M. Phillips, and R. Terlevich. “Classification parameters for the emission-line spectra of extragalactic objects.” In: 93 (Feb. 1981), pp. 5–19. DOI: 10.1086/130766.

[3] Michael L. Balogh et al. “Differential Galaxy Evolution in Cluster and Field Galaxies at z$\sim$0.3”. In: 527.1 (Dec. 1999), pp. 54–79. DOI: 10.1086/308056. arXiv: astro-ph/9906470 [astro-ph].

[4] Eric F. Bell et al. “Nearly 5000 Distant Early-Type Galaxies in COMBO-17: A Red Sequence and Its Evolution since z$\sim$1”. In: 608.2 (June 2004), pp. 752–767. DOI: 10.1086/420778. arXiv: astro-ph/0303394 [astro-ph].

[5] L. Binette et al. “Photoionization in elliptical galaxies by old stars.” In: 292 (Dec. 1994), pp. 13–19.

[6] J. Brinchmann et al. “The physical properties of star-forming galaxies in the low-redshift Universe”. In: 351.4 (July 2004), pp. 1151–1179. DOI: 10.1111/j.1365-2966.2004.07881.x. arXiv: astro-ph/0311060 [astro-ph].

[7] G. Bruzual A. “Spectral evolution of galaxies. I. Early-type systems.” In: 273 (Oct. 1983), pp. 105–127. DOI: 10.1086/161352.

[8] Nell Byler et al. “Nebular Continuum and Line Emission in Stellar Population Synthesis Models”. In: 840.1, 44 (May 2017), p. 44. DOI: 10.3847/1538-4357/aa6c66. arXiv: 1611.08305 [astro-ph.GA].

[9] Nell Byler et al. “Self-consistent Predictions for LIER-like Emission Lines from Post-AGB Stars”. In: 158.1, 2 (July 2019), p. 2. DOI: 10.3847/1538-3881/ab1b70. arXiv: 1904.10978 [astro-ph.GA].

[10] J. A. Cardelli, G. C. Clayton, and J. S. Mathis. “The relationship between IR, optical, and UV extinction.” In: *Interstellar Dust*. Ed. by Louis J. Allamandola and A. G. G. M. Tielens. Vol. 135. Dec. 1989, pp. 5–10.

[11] R. Cid Fernandes et al. “Alternative diagnostic diagrams and the ‘forgotten’ population of weak line galaxies in the SDSS”. In: 403.2 (Apr. 2010), pp. 1036–1053. DOI: 10.1111/j.1365-2966.2009.16185.x. arXiv: 0912.1643 [astro-ph.CO].

[12] H. Domínguez Sánchez et al. “Improving galaxy morphologies for SDSS with Deep Learning”. In: 476.3 (Feb. 2018), pp. 3661–3676. DOI: 10.1093/mnras/sty338. arXiv: 1711.05744 [astro-ph.GA].

[13] J. L. Donley et al. “Identifying Luminous Active Galactic Nuclei in Deep Surveys: Revised IRAC Selection Criteria”. In: 748.2, 142 (Apr. 2012), p. 142. DOI: 10.1088/0004-637X/748/2/142. arXiv: 1201.3899 [astro-ph.CO].

[14] C. P. Haines, A. Gargiulo, and P. Merluzzi. "The SDSS-GALEX viewpoint of the truncated red sequence in field environments at z ∼0". In: 385.3 (Apr. 2008), pp. 1201–1210. DOI: 10.1111/j.1365-2966.2008.12954.x. arXiv: 0707.2361 [astro-ph].

[15] T. M. Heckman. "An optical and radio survey of the nuclei of bright galaxies - Stellar populations and normal H II regions". In: 87.1-2 (July 1980), pp. 142–151.

[16] Luis C. Ho, Alexei V. Filippenko, and Wallace L. W. Sargent. "A Search for "Dwarf" Seyfert Nuclei. III. Spectroscopic Parameters and Properties of the Host Galaxies". In: 112.2 (Oct. 1997), pp. 315–390. DOI: 10.1086/313041. arXiv: astro-ph/9704107 [astro-ph].

[17] Luis C. Ho, Alexei V. Filippenko, and Wallace L. W. Sargent. "A Search for "Dwarf" Seyfert Nuclei. VI. Properties of Emission-Line Nuclei in Nearby Galaxies". In: 583.1 (Jan. 2003), pp. 159–177. DOI: 10.1086/345354. arXiv: astro-ph/0210048 [astro-ph].

[18] Guinevere Kauffmann et al. "The host galaxies of active galactic nuclei". In: 346.4 (Dec. 2003), pp. 1055–1077. DOI: 10.1111/j.1365-2966.2003.07154.x. arXiv: astro-ph/0304239 [astro-ph].

[19] L. J. Kewley et al. "Theoretical Modeling of Starburst Galaxies". In: 556.1 (July 2001), pp. 121–140. DOI: 10.1086/321545. arXiv: astro-ph/0106324 [astro-ph].

[20] Lisa J. Kewley et al. "The host galaxies and classification of active galactic nuclei". In: 372.3 (Nov. 2006), pp. 961–976. DOI: 10.1111/j.1365-2966.2006.10859.x. arXiv: astro-ph/0605681 [astro-ph].

[21] D. Christopher Martin et al. "The Galaxy Evolution Explorer: A Space Ultraviolet Survey Mission". In: 619.1 (Jan. 2005), pp. L1–L6. DOI: 10.1086/426387. arXiv: astro-ph/0411302 [astro-ph].

[22] S. Mateos et al. "Using the Bright Ultrahard XMM-Newton survey to define an IR selection of luminous AGN based on WISE colours". In: 426.4 (Nov. 2012), pp. 3271–3281. DOI: 10.1111/j.1365-2966.2012.21843.x. arXiv: 1208.2530 [astro-ph.CO].

[23] P. Papaderos et al. "Nebular emission and the Lyman continuum photon escape fraction in CALIFA early-type galaxies". In: 555, L1 (July 2013), p. L1. DOI: 10.1051/0004-6361/201321681. arXiv: 1306.2338 [astro-ph.CO].

[24] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *arXiv e-prints*, arXiv:1201.0490 (Jan. 2012), arXiv:1201.0490. arXiv: 1201.0490 [cs.LG].

[25] Clara M. Pennock et al. "Discovering exotic AGN behind the Magellanic Clouds". In: *Nuclear Activity in Galaxies Across Cosmic Time*. Ed. by Mirjana Pović et al. Vol. 356. Jan. 2021, pp. 335–338. DOI: 10.1017/S1743921320003270. arXiv: 2004.04531 [astro-ph.GA].

[26] Samir Salim et al. "GALEX-SDSS-WISE Legacy Catalog (GSWLC): Star Formation Rates, Stellar Masses, and Dust Attenuations of 700,000 Low-redshift Galaxies". In: 227.1, 2 (Nov. 2016), p. 2. DOI: 10.3847/0067-0049/227/1/2. arXiv: 1610.00712 [astro-ph.GA].

[27] Kevin Schawinski et al. "Observational evidence for AGN feedback in early-type galaxies". In: 382.4 (Dec. 2007), pp. 1415–1431. DOI: 10.1111/j.1365-2966.2007.12487.x. arXiv: 0709.3015 [astro-ph].

[28] David J. Schlegel, Douglas P. Finkbeiner, and Marc Davis. "Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds". In: 500.2 (June 1998), pp. 525–553. DOI: 10.1086/305772. arXiv: astro-ph/9710327 [astro-ph].

[29] R. Singh et al. "The nature of LINER galaxies:. Ubiquitous hot old stars and rare accreting black holes". In: 558, A43 (Oct. 2013), A43. DOI: 10.1051/0004-6361/201322062. arXiv: 1308.4271 [astro-ph.GA].

[30] M. F. Skrutskie et al. “The Two Micron All Sky Survey (2MASS)”. In: 131.2 (Feb. 2006), pp. 1163–1183. DOI: 10.1086/498708.

[31] Vasileios Stampoulis et al. “Multidimensional data-driven classification of emission-line galaxies”. In: 485.1 (May 2019), pp. 1085–1102. DOI: 10.1093/mnras/stz330. arXiv: 1802.01233 [astro-ph.GA].

[32] G. Stasińska et al. “Can retired galaxies mimic active galaxies? Clues from the Sloan Digital Sky Survey”. In: 391.1 (Nov. 2008), pp. L29–L33. DOI: 10.1111/j.1745-3933.2008.00550.x. arXiv: 0809.1341 [astro-ph].

[33] Christy A. Tremonti et al. “The Origin of the Mass-Metallicity Relation: Insights from 53,000 Star-forming Galaxies in the Sloan Digital Sky Survey”. In: 613.2 (Oct. 2004), pp. 898–913. DOI: 10.1086/423264. arXiv: astro-ph/0405537 [astro-ph].

[34] Edward L. Wright et al. “The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance”. In: 140.6 (Dec. 2010), pp. 1868–1881. DOI: 10.1088/0004-6256/140/6/1868. arXiv: 1008.0031 [astro-ph.IM].

[35] Donald G. York et al. “The Sloan Digital Sky Survey: Technical Summary”. In: 120.3 (Sept. 2000), pp. 1579–1587. DOI: 10.1086/301513. arXiv: astro-ph/0006396 [astro-ph].