ML based source classification

'Galactic activity diagnostics based on IR/optical photometry and ML methods'

INSTITUTE OF ASTROPHYSICS

RISE-CHASC Workshop
CfA, 2-3 August 2022

European Commission

Marie Skłodowska-Curie Actions

Elias Kyritsis

Lead collaborators: A. Zezas, C. Daoutis

Babis Daoutis

MSc student
University of Crete-Physics Department & Institute of Astrophysics

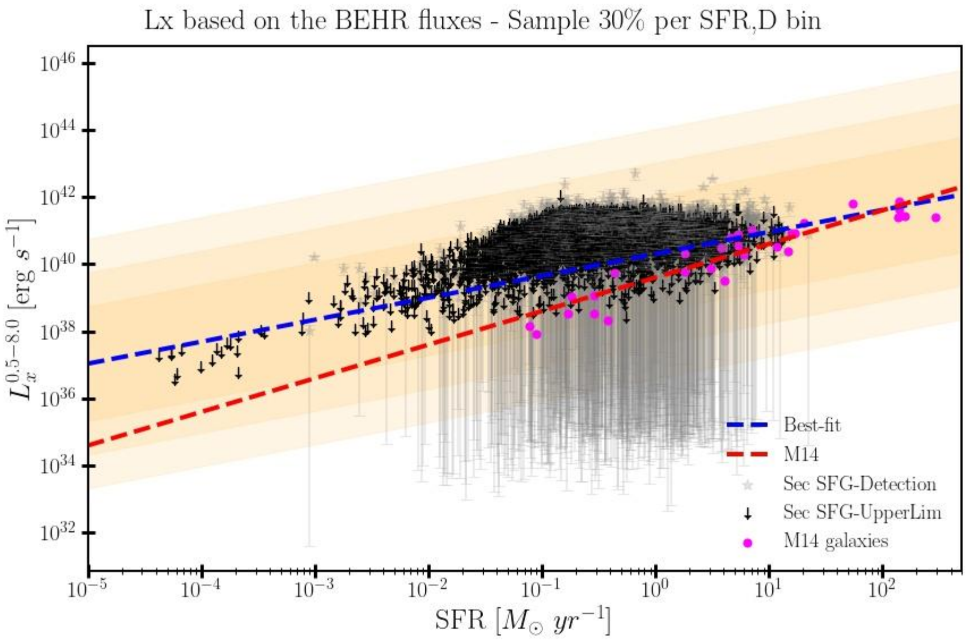
Motivation

Study the connection between X-ray luminosity of galaxies & their stellar population parameters (i.e. SFR, $M_\star$, Z)

I. Methodology for fitting unbiased scaling relations. ✓

II. What about the sample itself **?**
We need **well characterized** data .

The characterisation of a complete sample of *bona-fide* star-forming (or passive) galaxies is needed !



Traditional way of activity classification

1) Characteristic emission-line ratios - BPTs diagrams

Separate the galaxies into different classes depending on the source of ionization.

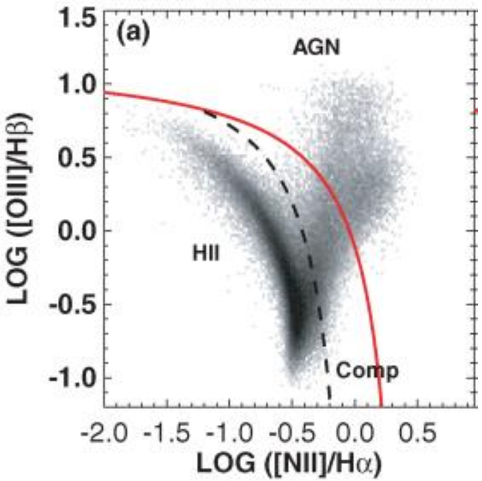
2) Stampoulis et al. 2019 developed a 4-D diagnostic following a soft clustering analysis.

Why do we need a new activity diagnostic ?

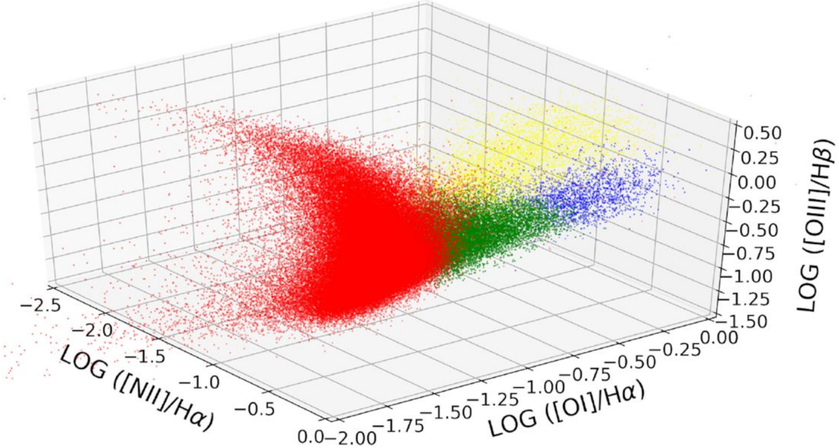
➢ The need of spectroscopic information limits the applicability of these diagnostics.

➢ Acquisition of more spectra is time expensive.

➢ Galaxies without emission lines cannot be classified.



[Kewley et al.,MNRAS,2006,372,961]



[Stampoulis et al.,MNRAS,2019,485,1085]

Traditional way of activity classification

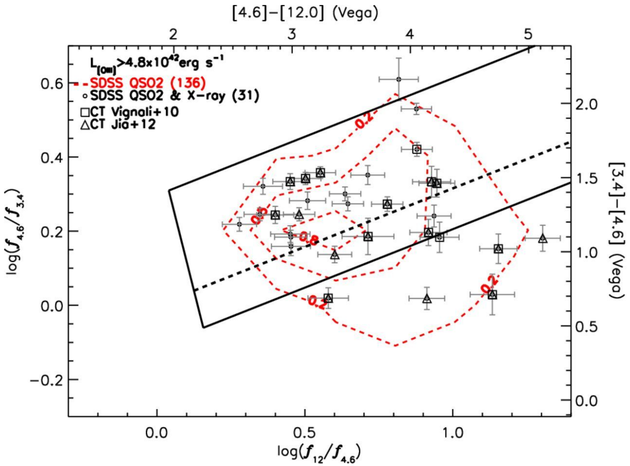
3) mid-IR/ multi-band photometry

- ➔ Widely use/ Well characterized
- ➔ Easily applied
- ➔ All-sky coverage (WISE)

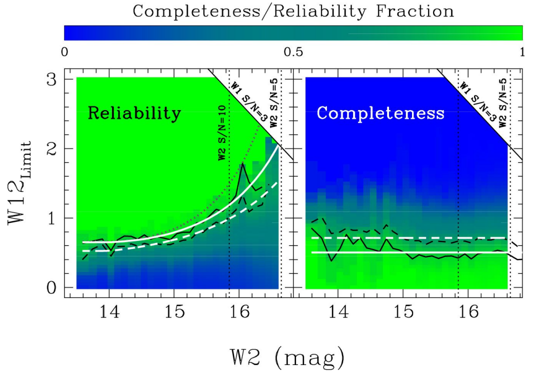
Why do we still need a new activity diagnostic ?

- ➢ Limited to identify only luminous AGN in high-redshift galaxies.
- ➢ Cannot discriminate the galaxies in other classes apart from star-forming and AGN.
- ➢ Not applicable in low redshift galaxies.

Development of a new galaxy activity classifier by training Machine Learning algorithm on multiwavelength data.



[Mateos et al.,MNRAS,2012,426,4,3271]



[Assef et al.,ApJ,2013,772,1,26]

Training sample

Definition of labels

Spectroscopic information:

SDSS-MPA-JHU catalog of galaxies

Applying Stampoulis et. al.,2019 to get the 4-activity classes.

- ➢ Using only spectra with S/N>5

Passive galaxies definition:

- ➢ Emission-line: S/N < 3 && Continuum: S/N > 3

5 Labels :

Star-forming, AGN, LINERs, Composite, Passive

Balancing the sample

z range: **0.02-0.08**
Strong imbalance between the classes as a function of z. (AGN & Passive galaxies dominate in high-z)

Splitting the training sample in 2 z bins: **low & high z .**

Balancing the sample according the number of objects per class in the low-z.

Total sample: 52001 galaxies

Class	Number of objects	Percentage (%)
Star forming	41425	79.7
Seyfert	2606	5.0
LINER	1640	3.1
Composite	3649	7.0
Passive	2681	5.2

Training sample

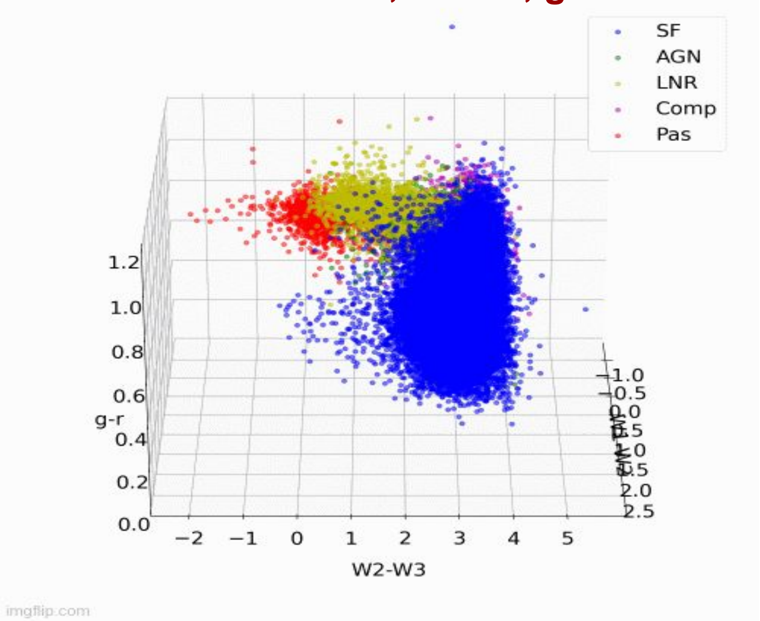
Definition of features

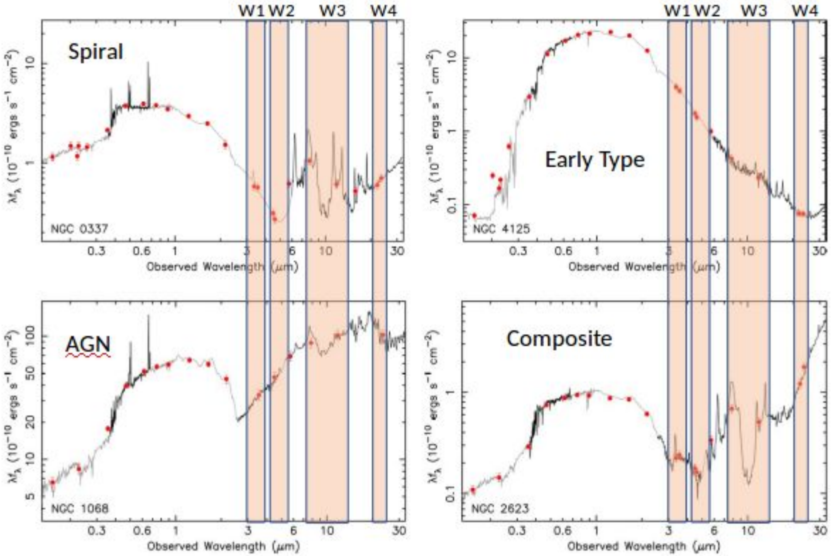
Photometric information:

- ➢ WISE all-sky survey: **W1,W2, W3 mid-IR bands**
- ➢ SDSS D16: **g,r** optical bands

3 Features

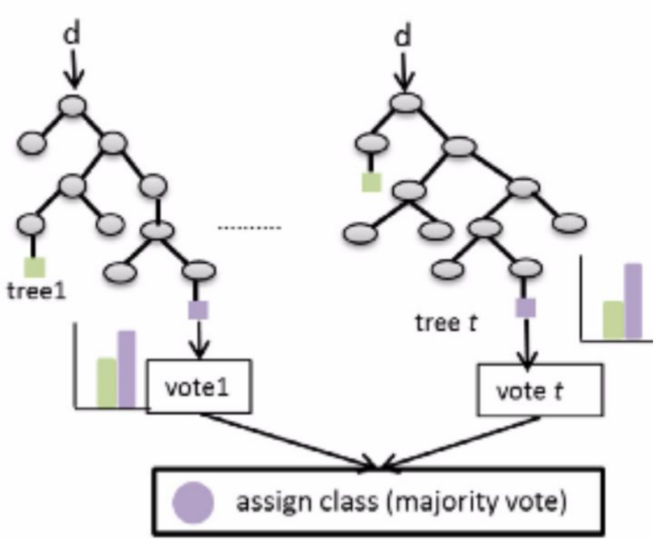
W1-W2,W2-W3, g-r





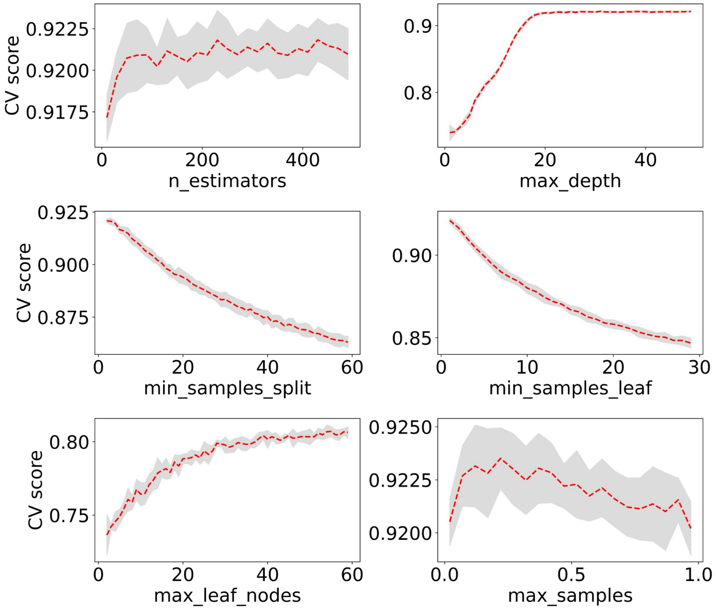
Random Forest algorithm and its Optimization

Random Forest Algorithm



Hyper-parameters tuning

Hyper-parameters tuning to reach the maximum performance of the algorithm



Based on the validation curves we defined a smaller range for the hyper-parameters within which a **GridSearch** was performed.

Final combination of Hyper-parameters values → **RF reaches the highest performance.**

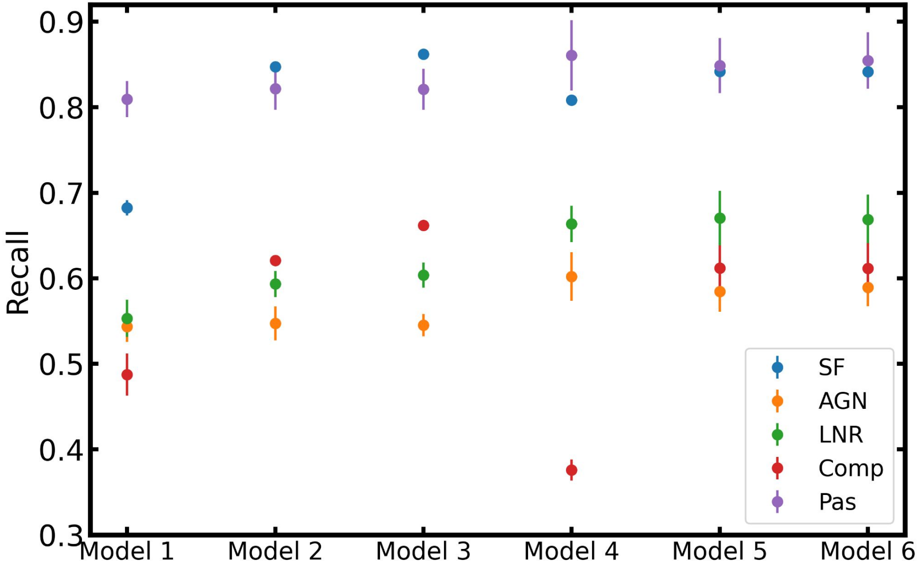
Random Forest algorithm and its Optimization

Feature optimization

Investigating if there is a specific combination of features that results in a better performance.

Evaluating the RF algorithm for different combinations of features.

I. Model 1: W1-W2, W2-W3
II. Model 2: W1-W2, W2-W3, g-r
III. Model 3: W1-W2, W2-W3,g-r, u-g
IV. Model 4: W1-W2, W2-W3, W3-W4
V. Model 5: W1-W2, W2-W3, W3-W4, g-r
VI. Model 6: W1-W2, W2-W3, W3-W4, g-r, u-g



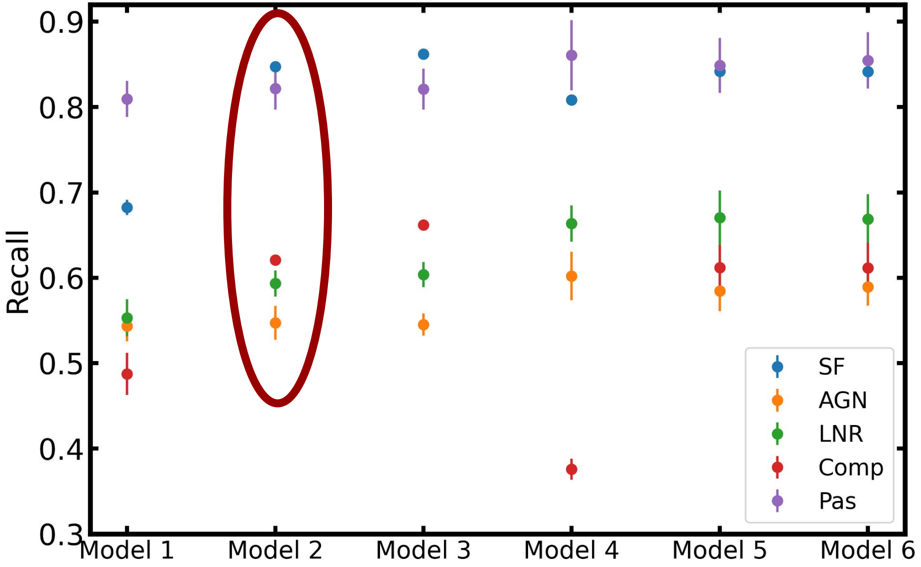
Random Forest algorithm and its Optimization

Feature optimization

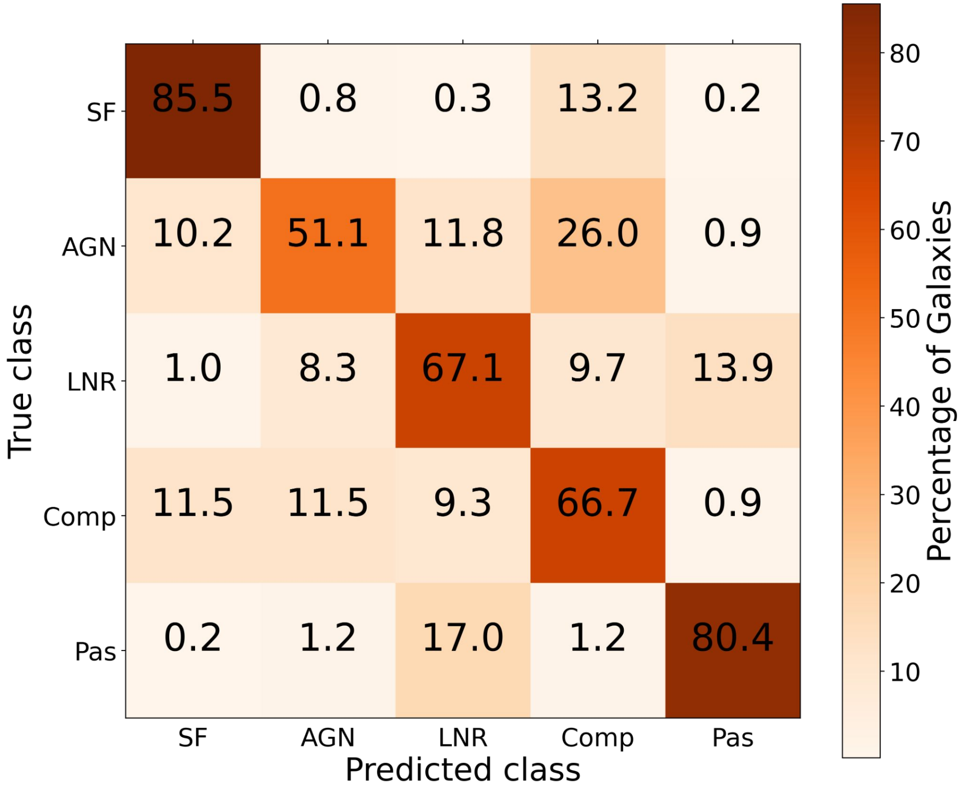
Investigating if there is a specific combination of features that results in a better performance.

Evaluating the RF algorithm for different combinations of features.

I. Model 1: W1-W2, W2-W3
II. Model 2: W1-W2, W2-W3, g-r
III. Model 3: W1-W2, W2-W3,g-r, u-g
IV. Model 4: W1-W2, W2-W3, W3-W4
V. Model 5: W1-W2, W2-W3, W3-W4, g-r
VI. Model 6: W1-W2, W2-W3, W3-W4, g-r, u-g



Results



Overall accuracy:

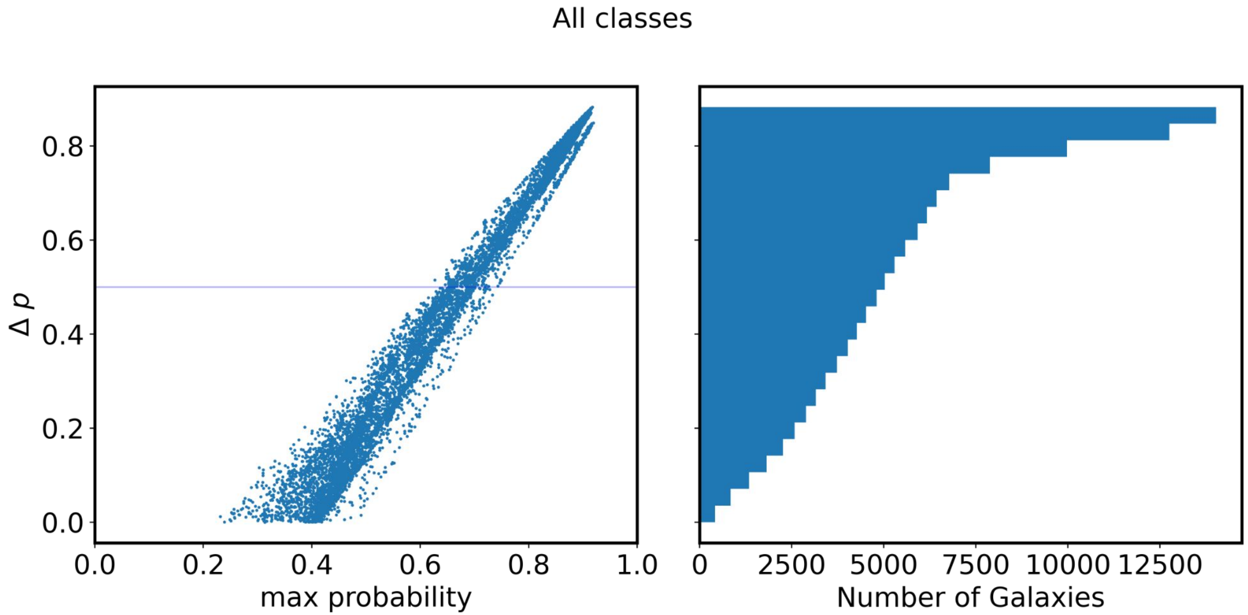
83% !

Best performing classes:
Star-forming & **Passive**

Reasonable characterization of:
LINERs & **Composite**

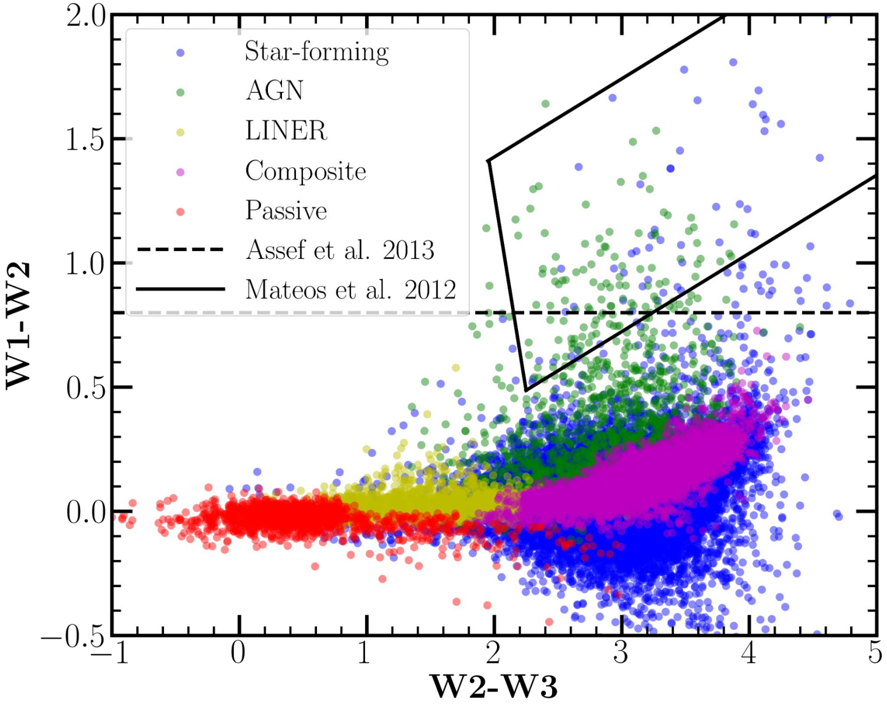
Poor characterization of : **AGN**

Results



Checking the confidence and the reliability of the algorithm
The results look very promising !

Application of the new diagnostic



Application of the activity diagnostic on the HECATE catalog

The classifier reveals a population of lower Luminosity AGN that the standard diagnostics cannot discriminate.

Take home message

★ A new activity diagnostic tool based on a RF classifier
  ○ No need for spectroscopic information.
  ○ Completely based on mid-IR and optical colors.
  ○ Applicable in large datasets and catalogs

★ Able to classify galaxies without emission lines.

★ High performance for Star-forming and Passive galaxies

★ High reliability and confidence on the predictions