ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

Galactic Spectra Classification using Cross-Correlation Methods

by

Errikos Papatheodorou

Submitted in Partial Fulfillment of the

Requirements for the Bachelor's Degree

in Physics

Supervised by Prof. Andreas Zezas

School of Sciences and Engineering
Department of Physics

University of Crete

Heraklion, Greece

October, 2024

Contents

1 Introduction **3**
1.1 Overview of Galactic Spectra Analysis . . . 3
1.2 Sample of Objects . . . 3
1.3 Importance of Spectral Classification . . . 4
1.4 Introduction to Cross-Correlation Method . . . 5

2 Data Collection and Preprocessing **5**
2.1 Data Acquisition . . . 5
2.2 Data Prepossessing & Cleaning . . . 7

3 Cross-Correlation Analysis **9**
3.1 Cross-Correlation Function . . . 9
3.2 Discrete Fourier Transform (DFT) . . . 9
3.3 Calculation of σ_a . . . 10
3.4 Deriving the R-value . . . 10
3.5 Classification uncertainty and measurement errors . . . 10
3.6 Classification . . . 11

4 Results **12**
4.1 Spectral Classification Accuracy . . . 12
4.2 Comparison with Other Classification Methods . . . 13

5 Discussion **14**
5.1 Interpretation of Results . . . 14
5.2 Overall assessment of the method . . . 15

6 Conclusion **15**
6.1 Summary . . . 15
6.2 Future Directions . . . 15

References **16**

1 Introduction

1.1 Overview of Galactic Spectra Analysis

Galactic spectra analysis serves as a cornerstone for investigating the complex mechanisms driving the activity within galaxies. By analyzing the light emitted across a range of wavelengths, we can gain crucial insights into the various excitation mechanisms responsible for the observed spectral features. These mechanisms, which include star formation, active galactic nuclei (AGN), and evolved stellar populations, interact with the gas and dust within galaxies, producing distinct spectral energy distributions. These distributions are reflected in the observed spectra, characterized by key features such as emission and absorption lines, which help us diagnose the nature of the radiation sources and the physical conditions of the gas.

The diagnostic power of galactic spectra lies in our ability to disentangle these excitation mechanisms, despite the complexity introduced by galaxies hosting multiple sources of ionization. A prime example is the Baldwin, Phillips, and Terlevich 1981 (BPT) diagram, which uses specific emission line ratios to categorize galaxies based on their predominant activity, distinguishing AGNs from star-forming galaxies and other intermediate types. However, such methods face limitations when dealing with composite galaxies, where both star formation and AGN activity coexist, or with passive galaxies, which lack strong ionizing radiation sources. Additionally, degeneracies related to metallicity and the intensity of the radiation field complicate the interpretation of spectra, often requiring complementary approaches like Spectral Energy Distribution fitting. This technique, while powerful, demands extensive and homogeneous datasets, restricting its broader applicability. Therefore, the challenge of accurately diagnosing galactic activity remains a central focus of modern galactic spectra analysis, as it provides a deeper understanding of the physical processes governing galaxy evolution.

1.2 Sample of Objects

- **Broad-Line Active Galactic Nucleus (BL-AGN)**

  These galaxies are characterized by broad emission lines in their spectra, indicative of fast-moving gas, often exceeding thousands of kilometers per second, near a supermassive black hole at the galaxy's center.

- **Narrow-Line Active Galactic Nucleus (NL-AGN)**

  In contrast to BL-AGNs, these galaxies display predominantly narrow emission lines, as the broad-line region of the spectrum is obscured by thick dust clouds. These dust clouds block the inner, high-velocity gas near the supermassive black hole, which typically produces broad spectral lines. As a result, only the narrow emission lines from the more distant, slower-moving gas remain visible in the observed spectrum.

- **Low-Ionization Nuclear Emission-Line Region (LINER)**

  LINERs are a class of galaxies where the spectral lines are dominated by low-ionization states of oxygen, nitrogen, and sulfur. These galaxies often feature weaker emission lines, suggesting a less intense ionizing radiation field, possibly from a low-luminosity AGN or old stellar populations.

- **Star-Forming (SF) Galaxies**

  The spectra of star-forming galaxies are dominated by strong emission lines from hydrogen, oxygen, and other elements, arising from regions of active star formation. These galaxies are typically rich in gas and dust, providing the raw material for new stars, and their spectra often exhibit the characteristics of young, hot stars.

  Blue Compact Dwarf (BCD) galaxies are a unique subclass of star-forming galaxies, characterized by their small size, low mass, and intense star formation activity. These galaxies often appear blue in optical images due to the presence of young, hot stars. Like other star-forming galaxies, their spectra exhibit strong emission lines, but BCDs stand out due to their particularly low metallicity. This indicates that they have undergone minimal chemical evolution, making them prime examples of galaxies in an early stage of star formation and evolution.

- **Composite Galaxies**

  These galaxies show a mix of spectral features, combining characteristics of both AGNs and star-forming regions. Their spectra often exhibit emission lines from both high-ionization and low-ionization states, suggesting a combination of processes, such as central black hole activity and widespread star formation.

- **Passive Galaxies**

  Passive galaxies, also known as quiescent galaxies, show little to no emission lines in their spectra, indicating a lack of recent star formation. Their spectra are dominated by the absorption lines of older stars, reflecting a population that has aged without significant new star formation activity.

This sample provides a comprehensive dataset that covers a wide range of galactic phenomena, from highly energetic AGNs to quiescent galaxies devoid of active star formation. These spectra will serve as the foundation for further analysis and classification.

1.3 Importance of Spectral Classification

Spectral classification is essential for unraveling the physical properties, formation histories, and evolutionary pathways of galaxies. By analyzing their spectra, we can extract key information, such as the age and distribution of stellar populations, star formation rates, and chemical abundances. This approach enables us to systematically categorize galaxies based on their spectral signatures, shedding light on the physical processes governing their development.

Moreover, spectral classification is particularly crucial in large-scale surveys, where vast amounts of spectral data need to be processed efficiently, making it an indispensable tool for modern astrophysical research. By enabling us to trace how galaxies evolve over cosmic time, spectral classification significantly contributes to our understanding of the universe's large-scale structure and dynamics.

1.4 Introduction to Cross-Correlation Method

In the context of galaxy spectra classification, cross-correlation serves as a powerful technique to compare the spectral data of galaxies with known templates or models. When classifying galaxies, we aim to group them based on their spectral features, which can reveal their stellar populations, star formation rates, and other intrinsic properties. However, due to noise and variations in observational data, direct classification can be challenging.

Cross-correlation addresses this by aligning the observed galaxy spectra with reference templates, which represent different galaxy types or classes. By systematically shifting the observed spectrum relative to the template and measuring how well they align, cross-correlation quantifies the similarity between the two. The best match provides a classification based on the closest corresponding template.

In galaxy classification, this technique allows for an automated approach, especially in large datasets like those from galaxy surveys. It ensures that even subtle features in the spectra, such as specific absorption or emission lines, are correctly identified and compared across different galaxy types, while at the same time providing an estimate of the galaxy's redshift.

2 Data Collection and Preprocessing

2.1 Data Acquisition

The dataset utilized in this study consists of spectra distributed across various classes, as detailed in Table 1. These spectra have been obtained from the Data Release 8 (DR8) of the Sloan Digital Sky Survey (SDSS), a comprehensive and widely utilized astronomical survey that provides high-quality spectral and photometric data for millions of celestial objects.

The dataset is divided into two categories: object spectra and template spectra, with an equivalent split across all classes. The template spectra are classified and serve as the reference set, while the object spectra, although labeled, have their labels hidden during the classification process and are only used afterward to evaluate the performance of the classification method. The distribution of data across these two categories is detailed in Table 1.

Table 1: Spectra distribution across classes and splits

Class	**Total Spectra**	**Test Spectra (20%)**	**Template Spectra (80%)**
SF	99	19	80
NL-AGN	99	19	80
BL-AGN	99	19	80
LINER	99	19	80
Composite	99	19	80
Passive	99	19	80
BCD	99	19	80
Total	**693**	**133**	**560**

In the classification of BL-AGN, the width of the hydrogen Balmer emission lines, is a critical criterion, typically measured by the full width at half maximum. These lines exhibit significant broadening due to the high-velocity motion of gas orbiting near the supermassive black hole.

The classification of the galaxies into their respective categories - NL-AGN, SF galaxies, LINER galaxies, and Composite galaxies — was performed using optical emission line ratios. This method, detailed in Stampoulis et al. 2019, leverages the BPT diagrams, which utilize the ratios of specific emission lines such as [OIII]/Hβ and [NII]/Hα to distinguish between different ionization sources.

To ensure the reliability of the classification, only emission lines with a signal-to-noise ratio (S/N) greater than 5 were considered. This threshold was applied to critical lines such as [OIII], Hα, Hβ, [NII], [SII], and [OI]. By enforcing this S/N criterion, the dataset minimizes uncertainties and ensures that the spectral features used for classification are robust and accurate.

Passive galaxies, characterized by their lack of recent star formation activity, were classified based on photometric properties, following the method outlined by Haines, Gargiulo, and Merluzzi 2008. This involved analyzing the colors and magnitudes of galaxies to identify those fitting the profile of quiescent, non-star-forming systems. These galaxies typically exhibit red colors and lack prominent emission lines, signifying an older stellar population that has ceased forming new stars. Our study utilizes a sample of passive galaxies from the dataset identified by Daoutis et al. 2023.

The BCD galaxies were identified using a combination of spectral and imaging criteria. Specifically, the D4000 continuum break index, a measure of the strength of the 4000 Å break in a galaxy's spectrum, was employed. BCD galaxies exhibit a D4000 index of less than 1, indicating a young, hot stellar population with minimal older stars. However, because D4000 < 1 is not a standard classification criterion for BCDs, a visual inspection was conducted to ensure that the selected galaxies exhibited flat spectra and compact, blue images in the SDSS data. This additional step confirmed the unresolved, starburst nature of BCDs, with ongoing star formation dominating their light output.

2.2 Data Prepossessing & Cleaning

The spectra obtained from the SDSS are provided with a logarithmic wavelength axis and are redshifted. To prepare these spectra for analysis using the cross-correlation tool, it is essential to preprocess and clean the data through several key steps.

Before performing the cross-correlation, the spectra must first be transformed from a logarithmic to a linear wavelength scale. In addition to delogging the wavelength axis, it is crucial to bring all spectra to a common rest frame by correcting for redshift. Since the galaxies in the sample have different redshifts, the spectra are deredshifted to remove the effects of their relative motion. This step aligns the spectral features across all samples, allowing for accurate comparison and analysis.

Division by the continuum is a crucial step in spectral analysis, particularly when performing cross-correlation. Cross-correlation is sensitive to the relative strengths of spectral features like absorption and emission lines, which are key indicators in comparing different spectra. However, variations in the continuum level—caused by factors such as differences in observational conditions or galaxy characteristics—can obscure these features and distort the analysis. By dividing the spectrum by its continuum, the continuum is flattened to a value 1, which eliminates differences due to the non-identical continuum levels of the spectra, focusing on the relative strength of the emission lines. This normalization allows the features to dominate the signal, ensuring that the cross-correlation focuses on the true variations in the spectra rather than noise or continuum discrepancies.

This normalization is performed by setting the DIVCONT to True in the file headers,which triggers the preprocessing pipeline to divide the spectrum by its continuum, ensuring consistent standardization. This process is critical for accurate cross-correlation, as it ensures that all spectra being compared are on the same continuum baseline, allowing the spectral features to be the primary contributors to the analysis.

Finally, spectra often exhibit increased noise levels towards their edges, where the signal-to-noise ratio (SNR) tends to diminish, as illustrated in Fig. 1a. To mitigate this issue, an edge-trimming algorithm is employed. This algorithm calculates the median SNR for a specified percentage of points at both the beginning and the end of the spectrum separately. Once the median SNR is determined, the algorithm searches for a predefined number of contiguous points where the SNR values exceed this median.

Upon identifying these contiguous points, the algorithm removes all preceding data points, effectively trimming the noisy edges, as demonstrated in Fig. 1c and Fig. 1d. This procedure shifts the focus of the analysis to the more reliable central regions of the spectrum, where noise is less prominent. By reducing the impact of edge noise, the algorithm enhances the quality of the data, improving the accuracy and reliability of the subsequent results. The edge-trimming step is crucial in the data pre-processing pipeline, as it ensures that low-SNR regions at the edges of the spectrum do not distort the cross-correlation function.

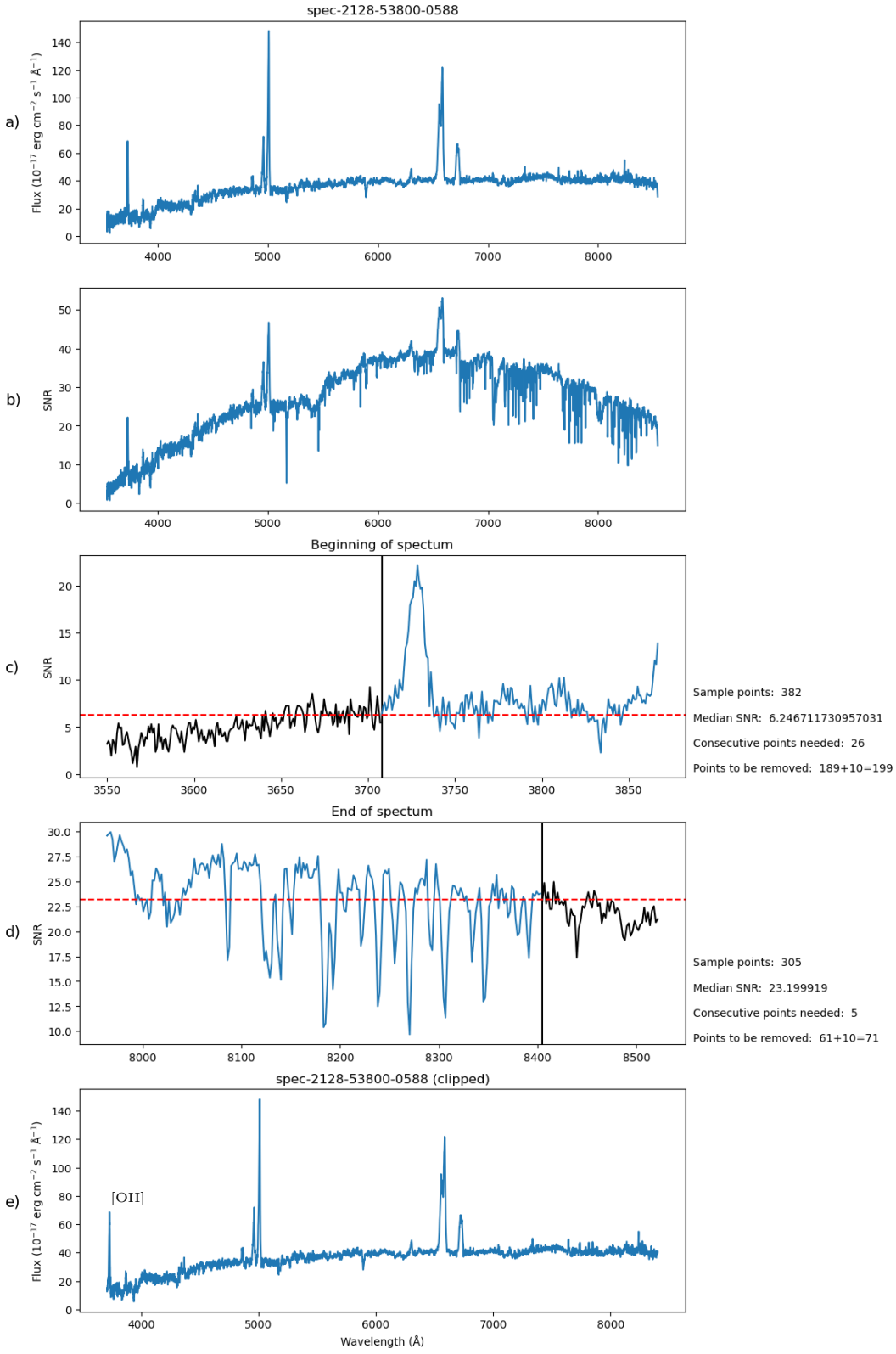

Figure 1: Example of the SNR clipping procedure: a) shows the initial spectrum, where edge noise is apparent. b) illustrates the corresponding SNR profile for this spectrum. In c) and d) all data points colored black are clipped. Finally, e) presents the resulting spectrum after clipping, with the noisy edge regions removed, while, in this example, the [OII] line was preserved.

3 Cross-Correlation Analysis

3.1 Cross-Correlation Function

The *cross-correlation function* (CCF) compares two discrete spectra, $g(n)$ (the object spectrum) and $t(n)$ (the template spectrum), to quantify their alignment at different shifts. The CCF is defined as:

$$c(n) = \frac{1}{N\,\sigma_g\,\sigma_t} \sum_{m=0}^{N-1} g(m)\,t(m-n) \tag{1}$$

where N is the number of discrete bins (samples), and σ_g and σ_t represent the root mean square (rms) values of the object and template spectra, respectively. The CCF reaches a peak at the shift n where the two spectra align best.

3.2 Discrete Fourier Transform (DFT)

Tonry and Davis 1979 chose to compute the cross-correlation function in Fourier space due to its computational efficiency, especially for large datasets. The galaxy spectrum $g(n)$ and template spectrum $t(n)$ can be transformed into Fourier space using DFT:

$$G(k) = \sum_{n=0}^{N-1} g(n)\,e^{-2\pi ink/N} \tag{2}$$

$$T(k) = \sum_{n=0}^{N-1} t(n)\,e^{-2\pi ink/N} \tag{3}$$

In Fourier space, the cross-correlation function simplifies to a product of the Fourier transform of the galaxy spectrum $G(k)$ and the complex conjugate of the Fourier transform of the template spectrum $T^*(k)$:

$$C(k) = \frac{G(k)\,T^*(k)}{\sigma_g\,\sigma_t} \tag{4}$$

To return to real space and retrieve the cross-correlation function $c(n)$, the inverse Fourier transform of $C(k)$ is applied:

$$c(n) = \frac{1}{N} \sum_{k=0}^{N-1} C(k)\,e^{2\pi ink/N} \tag{5}$$

Thus, the cross-correlation function can be obtained efficiently in real space after computing it in Fourier space.

3.3 Calculation of σ_a

To estimate the uncertainty in the cross-correlation function, we calculate the *rms noise* σ_a, which is derived by measuring the antisymmetric component of the cross-correlation function around its primary peak. The noise is given by:

$$\sigma_a = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left[c(n+\delta) - c(-n+\delta)\right]^2} \tag{6}$$

This formula captures the deviation from symmetry in the cross-correlation function about the main peak. For the proof of this formula, see Tonry and Davis 1979.

3.4 Deriving the R-value

The R-value is a measure of the signal strength in the cross-correlation function, defined as the ratio of the height of the primary peak to the level of noise. Let h be the height of the main peak in the cross-correlation function. The R-value is then calculated as:

$$R = \frac{h}{\sqrt{2}\,\sigma_a} \tag{7}$$

A higher R-value indicates a stronger correlation between the two spectra, as the primary peak stands out more prominently against the background noise. Conversely, lower R-values can signal that the object spectrum may not correlate as well with that specific template or that noise is present in the spectra.

3.5 Classification uncertainty and measurement errors

To gain insight into the uncertainty of our classification process, we conducted simulations to analyze the distribution of R-values associated with different templates. This way we can quantify whether the difference between the R-values of different templates is due to the noise in the data or indicates a clearly different template. To do this we introduced Gaussian noise into the object spectrum to simulate slight variations from the original data, resulting in 1000 simulated spectra. Then we performed the cross correlation analysis of each simulated spectrum against the same templates. This process generated a range of R-values, resulting in a Gaussian-like distribution for each template. Fig. 2 shows the histograms of the R-values for the top three ranking templates for one indicative spectrum. The primary objective of these simulations was to identify the three templates that produced the highest R-values, providing initial candidates for classification, while also proving information on the mixing between the templates.

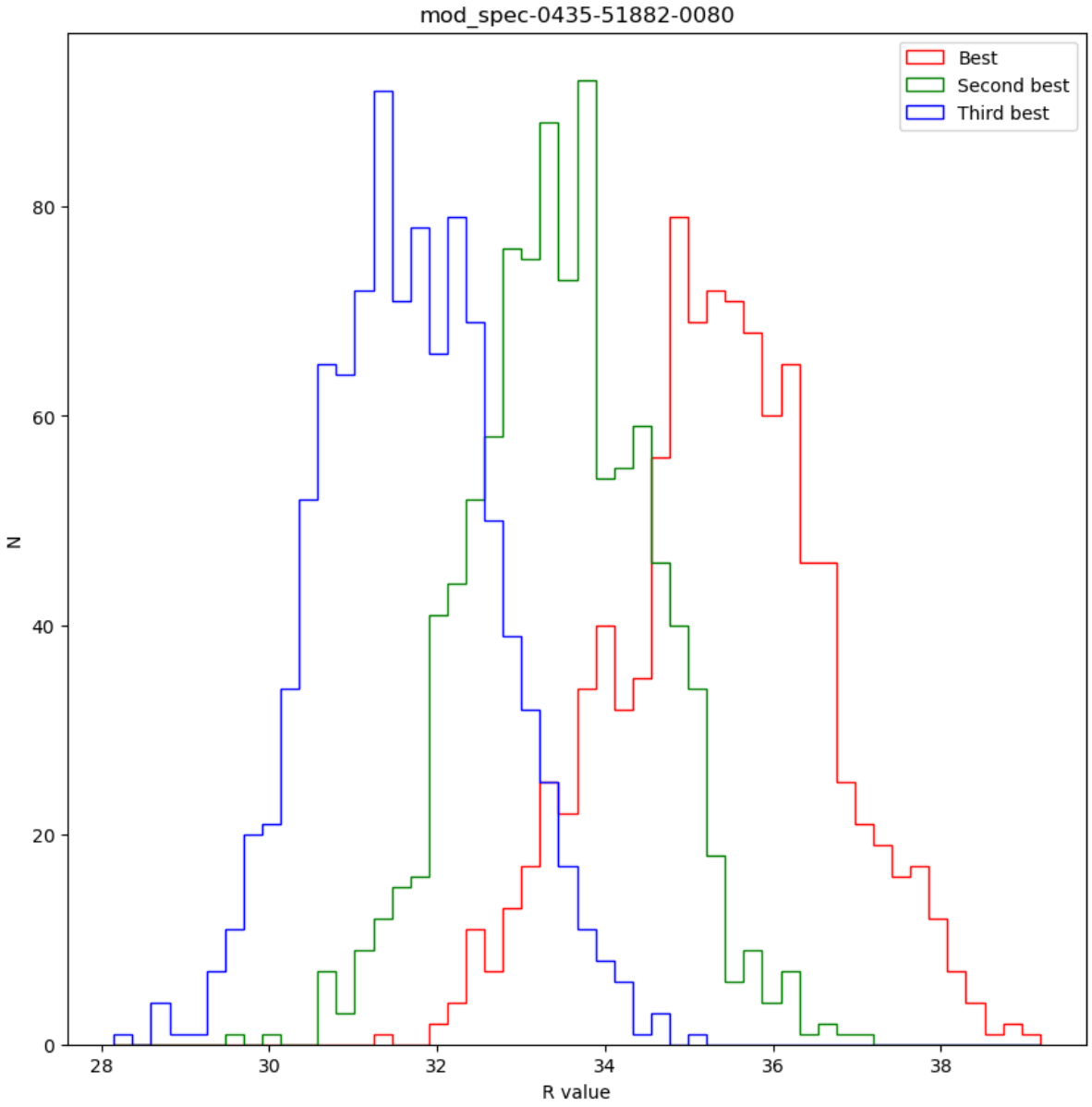

Figure 2: R-value distribution of best (red), second best (green) and third best (blue) template fits, where N is the number of object spectrum variants

The degree of overlap between the best-fit template and the second and third best templates provides a clear visual representation of the separation between templates. By analyzing this distribution, we can assess how distinctly the best-fit template stands out from the others, making it easier to evaluate the accuracy of template matching.

3.6 Classification

When we apply our method the reported R-value is the median of the distribution for each template. This way we limit statistical fluctuations and we account for uncertainties in the observed spectra. The class of the template spectrum with the highest median R-value was then selected as the predicted class for the object spectrum.

4 Results

4.1 Spectral Classification Accuracy

To calculate the accuracy of the cross-correlation method, we applied the classification procedure to the test set. This consists of 19 randomly selected objects from each class (20% of the original sample) as described in Section 2.1. To reduce computation time, when we assess the significance of the classification using the simulation method described in Section 3.5 we generated 100 varied versions of each object spectrum instead of the typical 1,000 variations, offering a balance between precision and efficiency.

To evaluate the overall performance, we constructed a confusion matrix, which compares the actual class of each spectrum to its predicted class. This matrix provided insight into both correct classifications and miss-classifications, allowing us to assess the accuracy of the method. Beyond this, we also calculated the recall for each class, measuring the model's ability to correctly identify instances of a particular class. Recall is the amount of correct classified spectra over the total spectra of each class. This metric allowed us to evaluate how well the model performed in recognizing individual classes, helping us identify potential areas where the model might struggle to correctly find certain spectra.

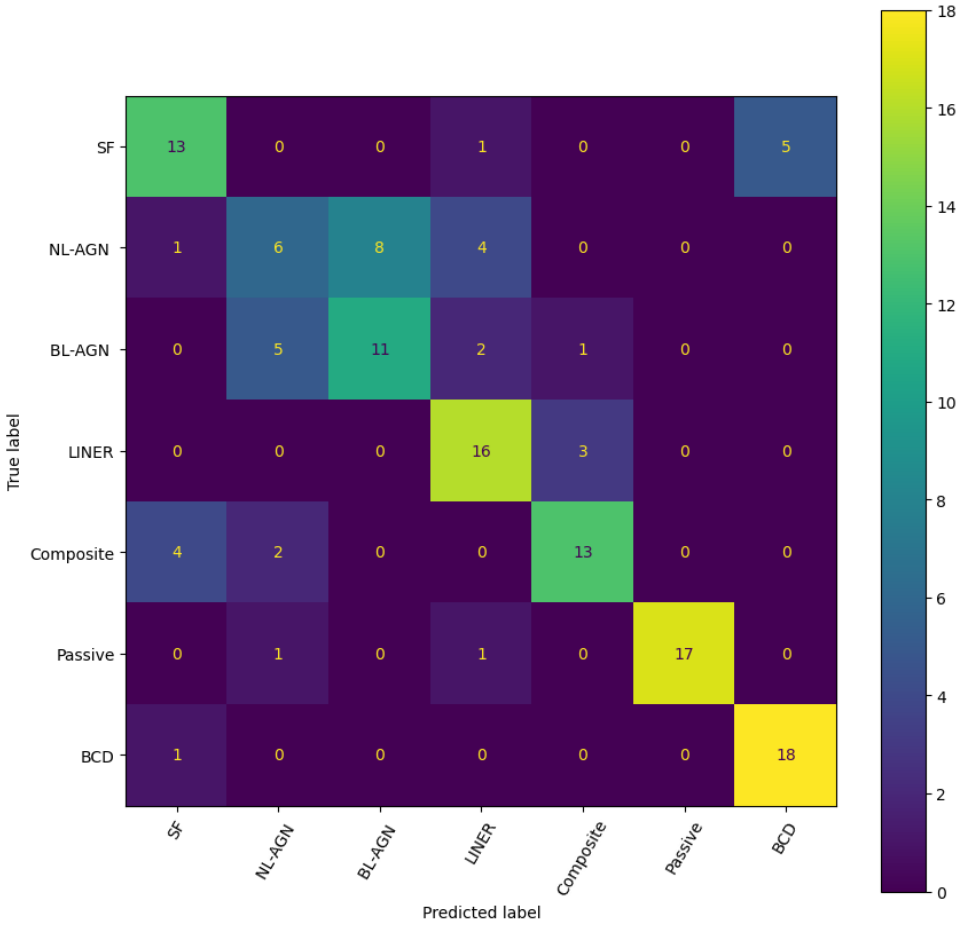

Figure 3: Confusion Matrix with the color bar representing the number of objects.

The confusion matrix, combined with the recall values and accuracy shown in Table 2, demonstrates that the method performs well in classifying BCD, Passive, and LINER galaxies.

Table 2: Recall values for each class and overall accuracy.

Class	**Recall**
SF	0.6842
NL-AGN	0.3158
BL-AGN	0.5789
LINER	0.8421
Composite	0.6842
Passive	0.8947
BCD	0.9474
Accuracy	**70.6767%**

The previous statement is further supported by recall values of 0.84, 0.89, and 0.95 for LINER, Passive and BCD galaxies respectively. However, it reveals challenges in accurately classifying NL-AGN and BL-AGN galaxies, with recall values of 0.32 and 0.58 respectively.

4.2 Comparison with Other Classification Methods

In order to compare with other methods (e.g. BPT diagrams or IR photometric methods) which do not discriminate between BL-AGN and NL-AGN or between SF and BCD (i.e. extreme SF) galaxies we have to group these classes. So by merging the SF and BCD classes into a single class named SF-all, and combining the BL-AGN and NL-AGN classes into one class named AGN, the recall for each individual class and the overall accuracy of the classification model naturally change. In Table 3 we present the recall scores for this simplified class scheme.

Table 3: Recall values for each class and overall accuracy with SF, BCD classes merged to SF-all and BL-AGN, NL-AGN merged to AGN.

Class	**Recall**
SF-all	0.9737
AGN	0.7895
LINER	0.8421
Composite	0.6842
Passive	0.8947
Accuracy	**84.9624%**

In their study, Daoutis et al. 2023, who used a classification method based on infrared and optical colors, achieved an accuracy of 82% using the same classification classes but on a significantly larger dataset. In contrast, our approach yielded an accuracy of 85%. While these results are promising, it is essential to emphasize that accuracy alone is not the sole metric for evaluating the

performance of a classification method. Our dataset, although smaller, was also more balanced, which can contribute to the robustness of our model's performance.

5 Discussion

5.1 Interpretation of Results

One of the major challenges we faced was the difficulty in accurately distinguishing between narrow and broad line AGNs. This was expected, as Fig. 4 clearly illustrates the similarity in the spectral features of both types, making their classification particularly challenging.

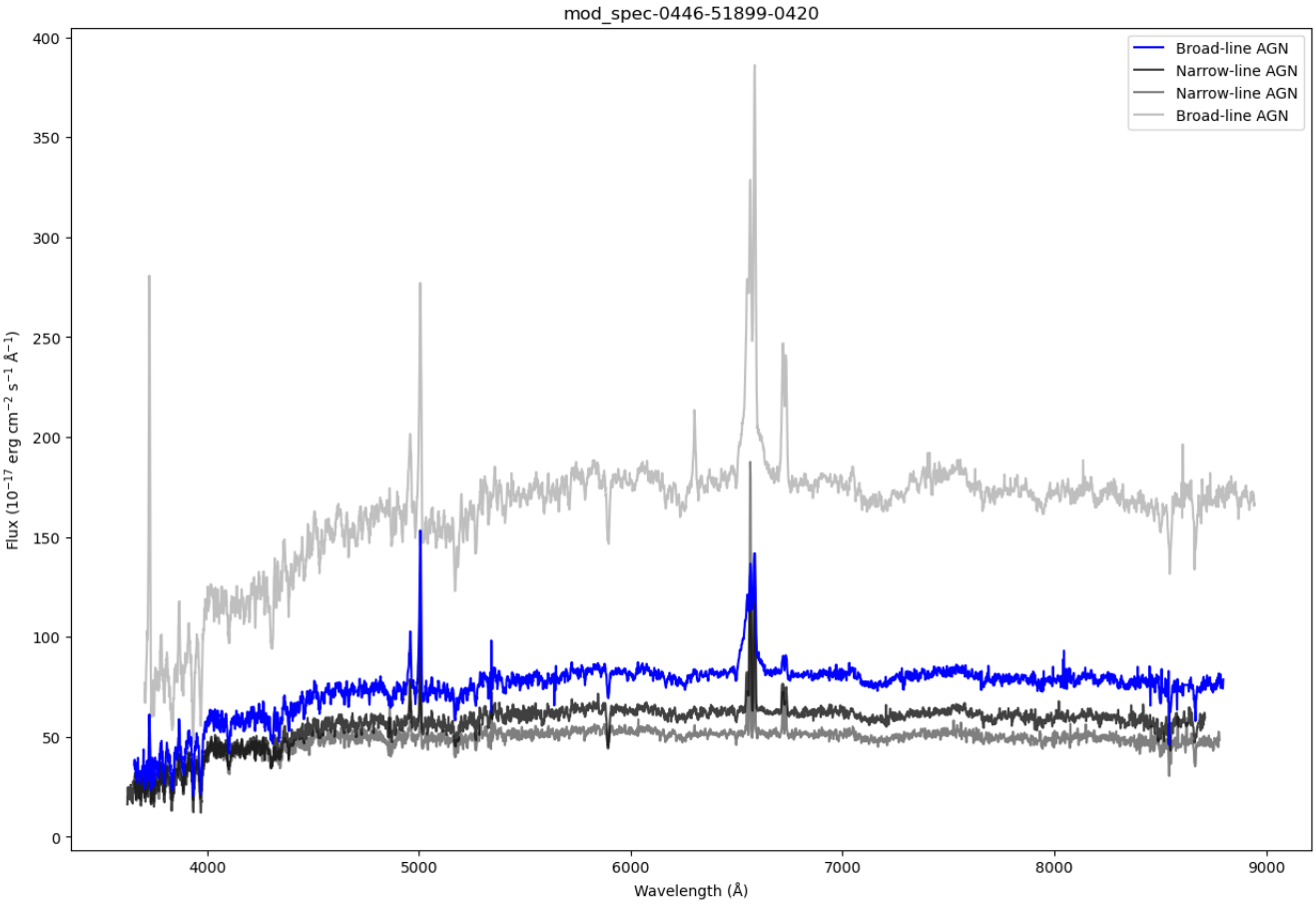

Figure 4: The blue line represents the object spectrum, while the black, dark gray, and gray lines show narrow-line and broad-line AGN templates in descending order of median R-value.

Overall the method works very well. The two classes that are mostly mixed are the BL-AGN and the NL-AGN. This is expected since broad-line AGN that have narrower lines may be confused in the cross-correlation analysis with the narrow-line AGN. Similarly BL-AGN that have very strong forbidden lines which are narrow may be mistakenly classified as narrow-line AGN based on the template matching. However, it is important that AGN are not significantly mixed with other classes. There only a small fraction of miss-classified AGN as LINERs which again could be explained on the basis of the similarity of the spectra of the two types of sources.

Regarding LINERs and Composites there is a small fraction of LINERs miss-classified as Composites and a small fraction of Composites miss-classified as SF or AGN, both of which are expected given the similarity of these spectra.

5.2 Overall assessment of the method

We have found that cross-correlation can serve as a valuable classification tool, particularly for galactic spectra with distinct spectral features. However, when dealing with classes that exhibit very similar spectral profiles—such as those that differ primarily in emission line widths—this method has some limitations, but the mixing between different activity classes is minimal. Furthermore, with the simulation analysis, we can assess the uncertainty in the classification due to uncertainties in the data.

6 Conclusion

6.1 Summary

The results of this study demonstrate that the cross-correlation method effectively classifies different galaxy types. It only faces minor limitations, particularly in distinguishing between galaxies with similar spectral features. The primary result, as shown in the confusion matrix, reveals an overall accuracy of 71%. The method performed well in classifying passive and BCD galaxies, with recall values of 0.89 and 0.95, respectively. However, it encountered significant challenges in accurately identifying NL-AGN and BL-AGN galaxies, which showed lower recall values due to the similarity in their spectral profiles. However, when we consider them as a single class (AGN) the accuracy in identifying AGN is 79%.

While cross-correlation methods have been generally used to classify broad classes of objects (e.g. stars, galaxies, QSOs etc) our analysis shows that with a correctly selected set of templates we can distinguish between different sub-classes of emission-line galaxies (SF, AGN) or even subclasses between them. An important aspect of our analysis is the possibility to assess the precision of the classifications while accounting for measurement errors.

6.2 Future Directions

Our future goals aim to strengthen and expand the validation of our method. First, we plan to apply the technique to a larger dataset to more robustly evaluate its effectiveness and ensure the results are statistically significant. Additionally, we want to explore how reducing the resolution of the spectra affects the outcomes. Moreover, an important next step will be to quantify the separation analysis of the templates by calculating the classification probability into a given class. This will provide a more objective measure of the method's performance, enabling a clearer interpretation of how well distinct classes are separated.

References

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

2. Daoutis, C. et al. (Nov. 2023). "A versatile classification tool for galactic activity using optical and infrared colors". In: 679, A76, A76. DOI: 10.1051/0004-6361/202347016. arXiv: 2310.02888 [astro-ph.GA].

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

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

5. Tonry, J. and Davis, M. (Oct. 1979). "A survey of galaxy redshifts. I. Data reduction techniques." In: 84, pp. 1511–1525. DOI: 10.1086/112569.