UNIVERSITÀ DEGLI STUDI DI ROMA TOR VERGATA
MACROAREA DI SCIENZE MATEMATICHE, FISICHE E NATURALI

CORSO DI STUDIO IN
FISICA - Curriculum Erasmus Mundus MASS

TESI DI LAUREA IN
Astrophysics and Space Science

TITOLO
From SEDs to Light-Curves: A Multimodal Approach to Broad Absorption Line Quasar Characterization and Identification

Relatore:
Prof. Francesco Tombesi, U. di Roma Tor Vergata

Correlatore:
Dr. Angela Bongiorno, INAF - OAR
Prof. Andjelka Kovačević, U. of Belgrade
Dr. Enrico Piconcelli, INAF - OAR

Laureando:
Matricola: 0325651
Nicolás G. Guerra Varas

Anno Accademico 2023/2024

MASTER IN ASTROPHYSICS AND SPACE SCIENCE

Erasmus Mundus Master
in Astrophysics and Space Science

Master Thesis

From SEDs to Light-Curves: A Multimodal Approach to Broad Absorption Line Quasar Characterization and Identification

Supervisors:

Dr. Angela Bongiorno

INAF - Osservatorio di Roma

Prof. Andjelka Kovačević

University of Belgrade

Author:

Nicolás G. Guerra Varas

Co-Supervisors:

Prof. Francesco Tombesi

Università di Roma Tor Vergata

Dr. Enrico Piconcelli

INAF - Osservatorio di Roma

Academic Year 2024/2025

TOR VERGATA UNIVERSITÀ DEGLI STUDI DI ROMA

UNIVERSITY OF BELGRADE

University of Bremen

UNIVERSITÉ CÔTE D'AZUR

Funded by the European Union

This Master thesis is submitted in partial fulfillment of the requirements for the degree FISICA - Curriculum ErasmusMundus as part of a multiple degree awarded in the framework of the Erasmus Mundus Joint Master in Astrophysics and Space Science – MASS jointly delivered by a Consortium of four Universities: Tor Vergata University of Rome, University of Belgrade, University of Bremen, and Université Cote d'Azur, regulated by the MASS Consortium Agreement and funded by the EU under the call ERASMUS-EDU-2021-PEX-EMJM-MOB.

Table of Contents

1. Introduction **3**
1.1. Active Galactic Nuclei . . . 3
1.1.1. What is an AGN? . . . 3
1.1.2. The Unification Model . . . 3
1.1.3. Multi-Wavelength Emission from AGN . . . 4
1.1.4. AGN Variability . . . 5
1.1.5. AGN Feedback . . . 5
1.2. Broad Absorption Line Quasars . . . 6
1.2.1. What is a BAL-QSO? . . . 6
1.2.2. BAL-QSO Spectroscopy . . . 7
1.2.3. BAL-QSO Variability . . . 8
1.2.4. AGN Feedback in BAL-QSOs . . . 9
1.3. Machine Learning in Astrophysics . . . 9
1.3.1. Multimodal Learning . . . 10
1.4. This Work . . . 11

2. Properties of the Reference Sample **13**
2.1. Presentation of the Sample . . . 13
2.2. Spectral Energy Distributions . . . 16
2.2.1. Multi-Wavelength Data . . . 16
2.2.2. Methods . . . 16
2.2.2.1. Corrections . . . 16
2.2.2.2. Obtaining the Mean SED . . . 18
2.2.2.3. Treatment of Errors . . . 18
2.2.3. Results . . . 18
2.2.3.1. SED for BAL QSOs vs. non-BAL QSOs . . . 19
2.2.3.2. SED for BALs in the Fully-in-g and Not-in-g Samples . . . 20
2.3. Spectra . . . 20
2.3.1. Composite Spectra . . . 21
2.3.1.1. Methods . . . 22
2.3.1.2. Results . . . 24
2.4. Light-Curves . . . 25
2.4.1. Methods . . . 28
2.4.1.1. Data . . . 28
2.4.1.2. Feature Extraction . . . 28
2.4.1.3. Comparison Tests . . . 30
2.4.2. Results . . . 31

2.4.2.1. Comparison of BAL and Non-BAL Features 31
2.4.2.2. Comparison of Fully-in-g BAL and non-Fully-in-g BAL Features 33
2.4.3. Future Prospects 34

3. Multimodal Learning Experiments **37**
3.1. Data Modalities 37
3.1.1. Spectra 37
3.1.2. Light-Curves 38
3.2. Training and Test Sample 38
3.3. Multimodal Learning Methods 40
3.3.1. Tree Ensemble Models 40
3.3.1.1. Spectral Dimensionality Reduction and Tabular Models . . 41
3.3.1.2. Light-Curve Tabular Models 43
3.3.1.3. Multimodal Random Forests 43
3.3.2. Dense Neural Network 47
3.3.2.1. Description of the Model and Fusion Technique 47
3.3.2.2. Results 48

4. Summary and Future Prospects **51**

References **55**

Abstract

Broad Absorption Line Quasars (BAL QSOs) are those that present strong absorption features in their spectra, which are associated to high-velocity outflows that go beyond 100 pc away from the central source, and can reach speed of $1.0c$. Thus, AGN feedback in these objects is particularly strong, as they are an ideal laboratory for studies on the effect of outflows on the host galaxy and its evolution. The Legacy Survey of Space and Time is expected to significantly increase the number of known BAL QSOs, allowing for novel, larger scale and ground-breaking AGN feedback studies. However, BAL QSO variability is not distinct from other QSOs. Therefore, their identification through light-curves is a challenge.

This thesis investigates the multi-wavelength properties and variability of BAL QSOs and explores the potential of multimodal machine learning for their identification in large-scale time-domain surveys such as LSST. We characterize a clean sample of 1419 BAL QSOs, derived from the SDSS DR7 QSO catalogue, and compare their properties to 41086 non-BAL QSOs. Our investigation includes the construction of mean Spectral Energy Distributions (SEDs), composite spectra, and light curve analysis. We use these to compare the BAL and non-BAL QSOs in our sample. We also compare those BAL QSOs which CIV absorption troughs land fully within the g-band of SDSS with the rest of the BAL QSOs, with the aim of testing whether the position of strong absorption features has a direct effect on the photometry of the object, and thus its SED and light-curves.

We recover the redder UV-to-optical continuum and steeper IR slope in BAL QSOs, associated with high dust extinction. Our results are consistent with previous works in the literature. Our composite spectra reveal the distinct characteristics of Hi-BALs, Lo-BALs, and FeLo-BALs. They also reveal that the shape, depth or blue-shift of CIV absorption does not necessarily correlate with the Balnicity index, which is used to define BAL QSO samples. We propose that studies dedicated to the details of the CIV absorption could potentially shed light on the details of the outflowing material in BAL QSOs, which can in turn provide insights on their relationship to the host galaxy and its evolution. Furthermore, we analyze the Zwicky Transient Facility (ZTF) light-curves by computing their time-domain features, and find no significant variability differences between BAL and non-BAL QSOs. We also find no significan differences in the SEDs nor variability between the BALs with CIV fully inside and outside the g-band of SDSS.

Moreover, to address the challenge of identifying BAL QSOs via variability, we develop multimodal machine learning models combining spectral and time-domain data. Our best-performing model, a dense neural network with multiplicative and attentive fusion, correctly identifies 74% of BAL QSOs in the test set, a significant improvement compared to 14% obtained by light-curve classification alone.

We note that the standard deviation of the lower subset defined by the Otsu thresholding algorithm was the feature found to have one of the most significant differences when statistically comparing light-curve feature distributions, and was also the most important

one for the classification done by the best performing tabular model, an extreme gradient descent tree ensembling. We propose further studies looking into this algorithm specifically to determine whether the result seen here is just a coincidence or bias, of if it can assist in BAL QSO classification through variability.

We emphasize the potential of multimodal learning approaches in astrophysics, and propose new ideas for potential implementations for the LSST, particularly in characterizing and identifying BAL QSOs for enabling future ground-breaking studies on their AGN feedback and galaxy evolution.

Chapter 1

Introduction

1.1. Active Galactic Nuclei

1.1.1. What is an AGN?

Galaxies whose central supermassive black hole (SMBH) (with a mass of $M_{BH} \sim 10^6 - 10^{11} M_\odot$) is accreting matter are denominated as "active", and active galactic nuclei (AGN) are defined as their central region. They are among the brightest sources in the Universe.

After the first unusual spectra observed by Fath [1909] and a couple serendipitous observations [Shields, 1999], there were no larger studies until the one by Seyfert [1943], were what are now known as Seyfert galaxies are described as nearby galaxies with unexpectedly broad lines found in their central region. Later, the 3C catalog [Edge et al., 1959] provided a larger sample of powerful radio sources that were later cross-matched with their optical conterparts, which appeared to be "point-like" or "quasi-stellar". These were called quasars (QSO), historically used to define higher luminosity AGN. The large redshift found by Schmidt [1963] for the 3C 273 QSO implied it is an extragalactic source, and thus that it is $\sim 10^{12}$ times more luminous than the Sun. At the same time, variability showed that the emitting region was about $\sim 1 - 10$ pc across. Stellar activity is not sufficient to explain such observations, and the theory for an accreting SMBH was developed by Salpeter [1964], and was later confirmed with X-ray observations [e.g. Elvis et al., 1978].

1.1.2. The Unification Model

The current picture for AGN consists of a unified model [de Lima Santos & Soltau, 2024, Netzer, 2015, Ramos Almeida & Ricci, 2017, Spinoglio & Fernández-Ontiveros, 2019] where the viewing angle explains the wide variety of observations seen in different kinds of AGN [Antonucci, 1993, Antonucci & Miller, 1985, Urry & Padovani, 1995]. This model is composed of a SMBH, an accretion disk, a corona, a dusty torus, a broad and narrow line region (BLR and NLR respectively), and a relativistic jet. Figure 1.1 shows an illustration of the unified model. This framework divides AGN into two broad categories: unobscured (type-I or Broad Line) and obscured (type-II or Narrow Line). The difference between them is the orientation angle with respect to the line of sight [Antonucci, 1993, Antonucci & Miller, 1985]: the former are seen at an angle with respect to the disk allowing the BLR to be observable, whilst the latter are seen edge-on, where the dusty torus covers these lines.

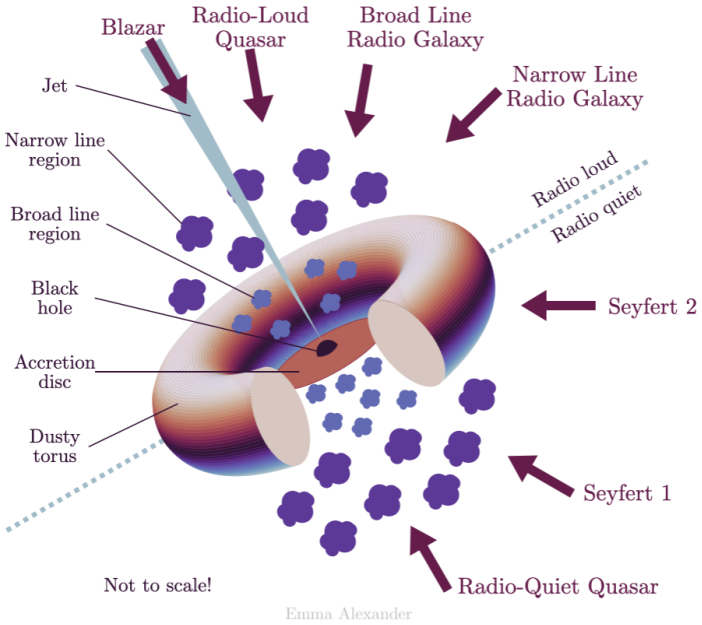


Figure 1.1: Illustration of the unified model of AGN by Alexander [Date of access: 2024].

1.1.3. Multi-Wavelength Emission from AGN

AGN can be detected accross the whole electromagnetic spectrum [Temple et al., 2021]. Their spectral energy distribution (SED) is a combination of many different physical processes and cannot be described as a single black-body (see Figure 1.2). In the ultra-violet (UV) to optical range, thermal emission from the accretion disk is observable and can be generally described as a broken power-law $L_\lambda \propto \lambda^{-\beta}$ or $L_\nu \propto \nu^{-\alpha}$ where β and α are the wavelength and frequency spectral indices respectively. In the near-infrared (NIR), emission from the hot dust in the torus dominates. Then, at higher energies, non-thermal emission from the corona dominates in the X-ray range. The BLR and NLR mostly emit between the UV and NIR ranges. The spectrum of an AGN has interesting elements at all wavelength ranges. It includes broad (Full Width Half Maximum FWHM $\geq$ 2000 km s^{-1}) and narrow emission lines, as well as both permitted and forbidden ones.

SED fitting has been crucial to AGN studies, as it can consicely provide insight into the underlying mechanisms of AGN activity as well as their host galaxy characteristics and their star-forming rates [e.g. Ciesla, L. et al., 2015, Marshall et al., 2022]. It has also played a key role in quantifying the contribution of the different components of AGN to the overall observed emission. For instance, Sokol et al. [2023] study the contribution from the torus and find that it is key to consider a varied enough range of models in order to avoid systematic biases and missed samples. Mountrichas et al. [2021] look into the X-ray contribution in SED modelling and find that it can have a significant role when separating the AGN and host contributions, and when deriving AGN properties. Overall, SED fitting is a key tool in AGN studies.

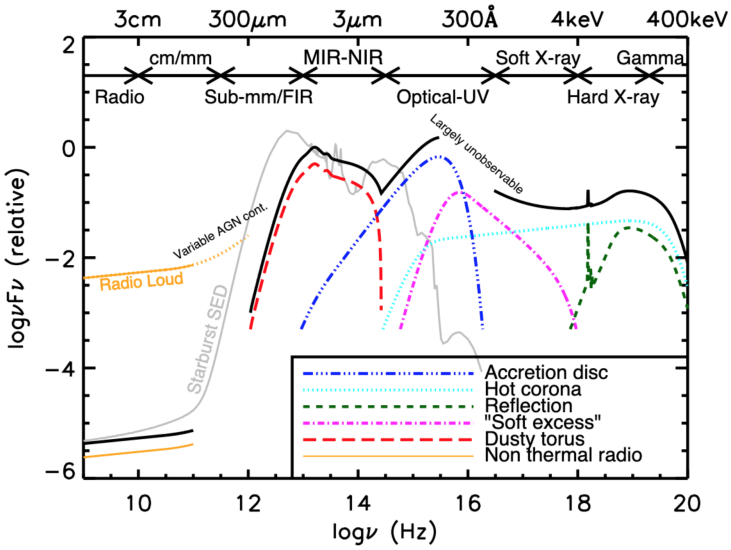


Figure 1.2: Characteristic AGN SED taken from Harrison [2014].

1.1.4. AGN Variability

AGN are variable across all bands [e.g. Hernández-García et al., 2016, Lira et al., 2015, Son et al., 2023, Ulrich et al., 1997] at time-scales of decades, years and as short as hours. Variations at shorter wavelengths are more rapid and originate from smaller and inner areas in the AGN structure. Some of the mechanisms responsible for the variations are instabilities in the disk, changes in the accretion rate or the distribution of obscuring material, or the influence of the relativistic jet on its surroundings. Variations have been detected in several ways such as changes in the continuum or overall brightness, as well as changes in the spectrum. For instance, Changing-Look AGN [e.g. Ricci & Trakhtenbrot, 2023] are those which degree of obscuration strongly vary, and/or present intermittent broad emission lines. Furthermore, different classes of AGN vary in distinct ways: type-I AGN are typically more variable, with shorter time-scales and larger amplitude; higher luminosity AGN or QSOs tend to show less extreme variability; and blazars (AGN whose jet is pointing in the line of sight) usually vary in the shortest time-scale and high amplitudes. Studying the variability behaviour of AGN is crucial for further understanding of these sources. The Zwicky Transient Facility (ZTF) [Masci et al., 2018] has played a key role in AGN variability studies. Given its high-cadence and sky coverage, it has provided an invaluably large dataset of AGN with long light-curves with many detections. The forthcoming Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST) [Kovacevic et al., 2021, LSST-Science-Collaboration et al., 2009, Sheng et al., 2022] will scale up the income of nightly data by at least an order of magnitude by continuous, deep and complete scans of the full sky, with improved cadence stategies. Studies of AGN variability have a promising future and are expected to improve with the contribution of the LSST.

1.1.5. AGN Feedback

Furthermore, the evolution of the host galaxies of AGN can potentially be hugely impacted by its activity, and even their surroundings and the SMBH itself [Fabian, 2012]. AGN are powerful enough to output radiation, winds, outflows and jets which can interact with the

inter-stellar medium (ISM) of the host. This could lead to galaxy quenching, which is when the star formation in the galaxy is haulted. AGN feedback can make the ISM far too hot for star formation, or even eject it from the host. Here, we overview a few relevant AGN feedback studies. By estimating the AGN coupling efficiency (i.e. the fraction AGN radiation that drives outflows from the host), Zubovas [2018] conclude that, as their title suggests, AGN are much more efficient at powering outflows as previously thought, and this could explain observations. Moreover, studies of AGN-driven outflows in low-redshift samples have revealed the impact they have on their host galaxies. For instance, Bessiere et al. [2024] found that, even though the star formation of the host galaxies in their QSO sample with AGN-driven outflows is not impacted within the studied timescale, they propose that multiple periods of AGN activity over a larger timescale could lead to the quenching of star formation. Torres-Papaqui et al. [2024] find evidence that two possible mechanisms of AGN outflow triggering are radiation and jets, and that the scenario where galaxies with more massive SMBH and larger bulges tend to exhaust the ISM in the host more rapidly and quench star formation plausible. Furthermore, Choi et al. [2020] ran cosmological simulations and found that AGN feedback is effective in ejecting metal-rich gas into the surrounding inter-galactic medium (IGM) of the host. Hopkins et al. [2016] found that AGN feedback can regulate the growth of the SMBH and can effectively mobilise obscuring material to a torus-like shape, consistent with observations. Overall, AGN feedback has a crucial impact on the host and studying it is a key part of galaxy evolution studies.

1.2. Broad Absorption Line Quasars

1.2.1. What is a BAL-QSO?

Broad Absorption Line Quasars (BAL QSOs) are those that present significant blueshifted absorption troughs in their spectra [Bischetti et al., 2023, Gibson et al., 2009, Hall et al., 2002, Lynds, 1967, Weymann et al., 1991]. BAL features are $\geq$ 2000 km s^{-1} wide, have velocities up to $0.1c$, or even $0.2c$ in some cases [e.g. Bruni et al., 2012, Rodríguez Hidalgo & Rankine, 2022]. These are associated with high-velocity outflows.

BAL QSOs are usually 10-20% of optically selected QSO populations [Gibson et al., 2009, Guo & Martini, 2019]. However, the fraction of BALs can increase up to $\geq$ 40% depending on the selection criteria of the sample. For instance, Maddox & Hewett [2008] find a larger BAL QSO fraction of $\sim$ 30% when selecting a sample based on IR colours. In addition, they have been found to be more prevalent at higher redshifts [Bischetti et al., 2022]. Furthermore, in spite of being a radio-quiet population in optically selected samples, BALs tend to appear at a higher fraction in radio-selected QSO samples [Bruni et al., 2019, de Kool, 1993, Menou et al., 2001, Petley et al., 2022]. They also tend to appear at a larger rate in high-luminosity samples [Bruni et al., 2019]. Indeed, Torres-Papaqui et al. [2024] found that wind velocity increases with luminosity.

Some studies have found indications that BAL QSOs are due to an orientation effect, whilst others postulate that they are rather a short evolutionary stage of intensified outflows in the early life of QSOs. In the first scenario, if the collimated outflow is along the line of sight, then the absorption features will be visible. Elvis [2000] proposes a unified structure for QSOs, and gives special attention to also explaining BAL features. He proposes that both BALs and narrow absorption lines (NALs) originate from the same outflow seen at different angles: the former appear when it is seen along the direction of the moving material, and the latter when

it is seen across the outflow. Supporting the same scenario, Lewis et al. [2003] find a trend of increasing flux with redshift in a small sample of seven BAL QSOs. They claim that since this trend is just as the ones seen in non-BAL QSO samples, the evolutionary scenario is not appropiate to explain this behaviour. In more recent years, Naddaf et al. [2023] assume the orientation-angle scenario and use a theoretical model with a radiation pressure mechanism for the acceleration of the dusty outflow to predict the probability of observing BAL features. They found that BAL effects increase with accretion rate and that BAL QSOs are slightly more massive than non-BAL QSOs ($> 10^8 M_\odot$), which is consistent with observational data from the Sloan Digital Sky Survey (SDSS), and thus supports the first scenario. Rankine et al. [2020] also support this scenario. They found the re-constructed CIV emission in BAL QSOs to be similar to that in non-BAL QSOs. On the other hand, Canalizo & Stockton [2002] support the idea of an evolutionary stage. In particular, they find that all four low-ionization BAL QSOs (Lo-BALs; see below) studied have a recently reactivated AGN and outflow activity triggered by mergers, and thus postulate that the BAL features are seen in an early-life evolutionary stage of QSOs. Furthermore, by analyzing the morphology of five BAL QSOs with very long baseline interferometry (VLBI) imaging, Montenegro-Montes et al. [2009] find a far too diverse geometry within the sample and indicate that this is difficult to explain with the orientation-angle approach. However, other studies have found that instead, a combination of both scenarios is required to fully explain the BAL phenomena [e.g. Nair & Vivek, 2022]. For instance, DiPompeo et al. [2013] found a significant IR dust excess in BAL QSOs and postulate that this can be explained by evolutionary models with a stage with a high-covering fraction, but note that the orientation angle most likely still plays an important role in the observed difference. Bruni et al. [2012] even considere a third possible scenario, where BAL features originate from polar jets accelerated by radiation pressure. They find that BAL and non-BAL QSOs are more of less the same age, ruling out the evolutionary scenario, and at the same time, that the spectra of the studied sample has a large range of spectral index, indicating a variety of orientation angles and thus ruling out the first scenario as well. Nowadays, this has not been finally resolved yet, though the orientation-angle scenario is generally more accepted.

BAL QSOs have been found to be more strongly reddened in the UV range than non-BAL QSOs, and more X-ray weak [Gallagher et al., 2007, Gibson et al., 2009, Saccheo et al., 2023]. Green et al. [2001] find that the modeled X-ray emission of BAL QSOs without the absorption effects due to the outflow are not intrinsically different from that in non-BAL QSOs, which reveals that the X-ray weakness is due to the absorbers.

1.2.2. BAL-QSO Spectroscopy

BAL QSO spectroscopy is the key aspect in this population. Several works have focused on modelling BAL QSO spectra by masking the absorption troughs and reconstructing the unabsorbed emission. Brodzeller & Dawson [2022] do this with a Principal Component Analysis (PCA) representation of the spectra to model a broad variety of QSO spectra in SDSS, and Rankine et al. [2020] are able to recover the unabsorbed UV emission and reconstruct the details of the spectra. They also build composite spectra to find general trends in the absorption and emission characteristics of the BAL QSOs. Maddox & Hewett [2008] also study the composite spectra divided by IR colours, and Mas-Ribas & Mauland [2019] find signatures of radiation pressure as the main mechanism for outflow acceleration in their composites.

The BAL QSOs are divided into three categories according to the ionization level of their

absorption lines [Gibson et al., 2009, Naddaf et al., 2023, Trump et al., 2006]:

1. Hi-BALs: These QSOs present BALs in a high-ionization state. Unambiguous identification includes verification that there are no BALs in the AlIII and MgII regions.

2. Lo-BALs: These present BALs in a low-ionization state, specifically in the MgII and AlIII regions. They might also be present in the SiIV and CIV regions. About 1.3% of QSOs are part of this class.

3. FeLoBALs: About 0.3% of QSOs present absorption from FeII as well as Lo-BAL features. These BALs could even be have specific evolutionary stage. Leighly et al. [2024] found two groups of FeLoBALs, associated with high and low accretion rates, and propose that these classes are part of an evolutionary sequence where the torus in FeLoBALs with low accretion rates cannot sustain optically thick winds.

4. MiniBALs: These QSOs present troughts that are associated to blended narrow-line absorption rather than BALs. So, these are not true BAL QSOs and contaminate BAL samples Hamann et al. [2013].

A rigorous definition for BALs was introduced by Weymann et al. [1991]. A QSO will be labeled as a BAL QSO if its Balnicity Index (BI) is positive:

$$\mathrm{BI} \equiv \int_{3000}^{25000} \left(1 - \frac{f(V)}{0.9}\right) C dV, \tag{1.1}$$

where $f(V)$ is the continuum-normalized spectral flux as function of rest-frame velocity V in km s^{-1}. The constant C is set to one if $f(V) < 0.9$, i.e. the expression in brackets is positive, for an interval of 2000 km s^{-1} or more.

Later, Hall et al. [2002] proposed the Absorption Index (AI) as a less restrictive measure capable of gaging the strength of all absorption features, not just broad ones:

$$\mathrm{AI} \equiv \int_{0}^{25000} \left(1 - \frac{f(V)}{0.9}\right) C dV \tag{1.2}$$

Here, $C = 1$ for when $f(V) < 0.9$ for an interval of al least 450 km s^{-1}.

1.2.3. BAL-QSO Variability

Furthermore, BAL QSOs are thought not to have a characteristic variability behaviour that can easily distinguish them from other QSOs. Type-2 AGN in general are less variable due to their obscuration [De Cicco et al., 2022], and it is possible that the outflows obscuring the central source play a similar role. Sánchez-Sáez et al. [2018] found that the structure function of BAL QSO variability is not distinct from other AGN. However, BAL features themselves have been found to be strongly variable. In particular, several studies have focused on the variability of CIV absorption troughs and have found that it can present drastic varitability in timescales as short as tens of hours and as long as years [De Cicco et al., 2017, Erakuman & Filiz Ak, 2017, Gibson et al., 2008, Green et al., 2023, Robinson et al., 2024]. This variability can indicate changes in the distribution and dynamics of the absorbing material. Capellupo et al. [2013, 2011, 2012] conduct an integral study of the CIV trough variability of 24 closely monitored BAL QSOs across short and long timescales (0.02

to 8.7 years in the rest-frame). They found that variability is more likely to be observed at longer timescales, suggesting that the changes in the outflowing material that give origin to the variability do not occur in the immediate surroundings of the central source (see also Welling et al. [2014]). They also find that stronger variability tends to be observed in weaker BAL features (see also Lundgren et al. [2007]). Additionally, Ruan et al. [2016] find that their QSO sample selected via variability has a larger fraction of BAL QSOs than samples selected and hypothesise that the variability in their spectra can subsequently have an effect on photometry and thus light-curves. So far, methods for the identification of BAL QSO samples using variability only remain to be developed.

1.2.4. AGN Feedback in BAL-QSOs

Moreover, the same outflowing material in BAL QSOs that originates variability can power AGN feedback processes, which tend to be particularly strong (e.g. kinetic luminosities larger than 10^{-3} times the bolometric luminosity of the QSO) [McGraw et al., 2017, Miller et al., 2020]. Several studies of individual BAL QSOs [e.g. Arav et al., 2013, Chamberlain et al., 2015] as well as larger works have found that these outflows lie more than 100 pc away from the central source (i.e. beyond the NLR and dusty torus), which is in agreement with BAL variability studies. The AGN feedback found in BAL QSOs could have a key role in the evolution of galaxies and their central SMBH across cosmic time. By studying BAL features in QSOs until redshift $z \sim 6$, Bischetti et al. [2023] find that in the early Universe (1 Gyr old) feedback from BAL QSO outflows is already ocurring, and additionally, that at larger redshifts BAL QSOs tend to be more prevalent and have stronger outflows. Bischetti et al. [2022] study this effect in high-luminosity QSOs. Within the evolutionary scenario of BAL QSOs, they propose that the BAL phase slows down SMBH and host growth, explaining the tendency for more massive galaxies at larger redshifts. Indeed, BAL QSOs are the ideal laboratory for AGN feedback studies.

1.3. Machine Learning in Astrophysics

As early as the nineties [Odewahn, 1998], astronomers have been interested in the potential that Machine Learning (ML) has to offer for the automization of time-consuming tasks, and since then, the interest and applications have grown up to 10,000 published astronomy papers that mention ML in their abstract [Borne, 2009, Djorgovski et al., 2022, Saraswat & Jain, 2021, Smith & Geach, 2023, Webb & Goode, 2023]. It is currently a vast field with many key works with useful models implemented. ML in general can be divided into supervised (when the data is accompanied by the correct output) and unsupervised (when it is not). Supervised ML is useful for tasks such as classification (to identify to which category data instances belong to), regression (to predict a value) or object detection, whilst unsupervised ML can assist in anomaly detection (identifying unusual or unexpected objects), dimensionality reduction (reducing the number of variables under consideration) or clustering (grouping similar data points together).

ML has been particularly useful in assisting AGN variability studies and enabling discoveries otherwise not plausible. In particular, ZTF and LSST brokers have implemented systems that allow for almost real-time monitoring of the varying sky [Förster et al., 2021, Möller et al., 2020, Matheson et al., 2021, Nordin, J. et al., 2019]. ZTF and the future LSST output alerts every night when variability of any kind is detected through these brokers, which can

enable quick follow-up observations such as obtaining spectroscopic data in early stages of supernova transients [Pruzhinskaya et al., 2023]. Sánchez-Sáez et al. [2021b] presents the light-curve classifies used by the ALeRCE broker[1]. Savić et al. [2023] also implement classifications for AGN light-curves. These, among other studies, have found that representing light-curves with time-domain features provides the best possible classification performance. De Cicco et al. [2021] estimate there to be 6.2×10^6 AGN in the LSST, proving how relevant it is that there are sufficiently efficient models in place to process the sheer volumes of expected data.

Moreover, Sánchez-Sáez et al. [2021a] and Baron & Poznanski [2017] have implemented anomaly detection algorithms capable of finding changing-look AGN via variability, and unusual galaxy spectra. Spectroscopic dimensionality reduction, classification and redshift estimation has been the focus of many relevant works such as Portillo et al. [2020], Iwasaki et al. [2023] and Szakacs et al. [2023]. Unsupervised methods, such as clustering of spectra [Teimoorinia et al., 2022] or building complete AGN samples [Hviding et al., 2024], just to name a few, have also been implemented and have prooved useful.

Models applied specifically to BAL QSOs have been developed as well, mostly focusing on spectral processing. Kao et al. [2024] test which dimensionality reduction techniques can be the most useful at providing useful spectral representations for classification. For BAL QSO identification in larger QSO samples, Nair & Vivek [2022], Guo & Martini [2019] and Busca & Balland [2018] build different convolutional neural networks (CNNs), Reichard et al. [2003] use tree ensemble models instead, and Tammour et al. [2016] use an unsupervised algorithm to cluster SDSS spectra. All these approaches have proved useful and are effective at assisting discoveries and insights into the open questions regarding BAL QSOs.

1.3.1. Multimodal Learning

Furthermore, in the past years, there has been a significant increase of multimodal machine learning (MML) models in the ML field [Akkus et al., 2022, Baltrušaitis et al., 2017, Liang et al., 2023, Ngiam et al., 2001, Parcalabescu et al., 2021, Sleeman IV et al., 2021]. To understand these models, the concept of "modality" should be defined first: a modality refers to the way in which something exists or is done. Then, a multimodal model can process and relate information from heterogeneous and interconnected modalities.

When building a MML model, one of the most important choices to make is how to fuse or ensemble the modalities [Liu et al., 2018, Sleeman IV et al., 2021, Zhao et al., 2024]. Different fusion techniques have diverse advantages and disadvantages, and should be chosen to best accommodate the specific data types and ML task. Additive fusion can be devided into early, intermediate and late fusion. Early fusion consists of creating a joint representation of the modalities, which can potentially be meaningful, and training a single model on it. However, it is tends to require extensive pre-processing of the individual modalities, such as specialized feature extraction, in order to combine them in a consistent way, and runs under the assumption that the model trained on the joint representation of the data is well suited for all modalitites. In late fusion, a separate model is trained on each data modality, and are fused at the decision level by, for instance, averaging or voting over the predictions of each model. This allows for each model to be specialized on the given data modality. However, it struggles to learn deeper correlations between the modalities. Overall, additive fusion techniques have the key disadvantage of having the implicit assumption that all the

[1] See also https://alerce.online/

modalities are equally reliable. Multiplicative and attentive fusion can be an alternative. The former consists of combining the modalities through multiplications like tensor products rather than in an additive way [Mittal et al., 2019]. It is good at capturing more complex interactions between the modalities. However, it can be more computationally expensive than other methods. Attentive fusion refers to the application of attention mechanisms [Vaswani et al., 2023] to give more relevance to a certain modality by assigning weights to, for instance, the decision probabilities. Even though it might require significant tuning, this technique is particularly useful when not all modalities are equally reliable.

So far, applications of this approach have mostly been in the industry, in domains such as: medicine, for the combination of medical images, patient records and other data; autonomous vehicles, for the combination of data from cameras, LiDARs, radars and other inputs; stock price prediction, for the combination of price time-series, tabular data, graphs and text from news articles and such; emotion recognition, for the combination of tone of voice, facial expressions and body language; and robotics, to allow robots to interact with multiple aspects of their environment.

This approach has made its way to the field of astrophysics, doubling the amount of papers with the keyword "multimodal" in the title and/or abstract in 20 years. Results are promising, showing the potential MML has for advancing the applications of ML to astronomical data. Cuoco et al. [2021] propose a MML model for multi-messenger astronomy, where data from joint γ-ray bursts and gravitational waves are combined in order to characterize astrophysical events. They apply this approach to combine not only two modalities (images and time-series), but also data from different instruments. Liu et al. [2023] the late multimodal fusion technique to combine a series of CNNs, each trained with a specific type of pulsar diagnostic image. For the estimation of photometric redshifts of QSOs, Hong et al. [2023] design a model with two main components: the first is capable of generating spectral features from photometric data; and the second applies multimodal transfer learning (i.e. the use of the knowledge gained from one task for improving performance of a second related task) by using the generated spectral features to increase the accuracy of photometric redshift prediction. Alegre et al. [2024] combine images and tabular data in order to build a MML model that can effectively distinguish single-component from multi-component radio sources in LOFAR data and obtain high accuracies. Parker et al. [2024] build an ensemble of two self-supervised transformers especialized on galaxy spectra and galaxy images each, with which a latent space of meaningful representations is created. The image and spectrum of a galaxy to be close in the latent space, which results in high performance for the tasks of galaxy morphology classification, photometric redshift estimation and galaxy property prediction.

Furthermore, even some multimodal Large Language Models (LLMs) specialized in astrophysics have been built. Mishra-Sharma et al. [2024] build PAPERCLIP, a model which connects astronomical images with natural language abstracts from successful observing proposals, by associating elements in the images to keywords in the abstracts.

These innovative models show promising results from applying a multimodal approach to astronomical data. MML models will most likely continue to be built for astropysics and assist advances in the field.

1.4. This Work

With forthcoming big data surveys in mind, such as the LSST [LSST-Science-Collaboration et al., 2009], which will provide up to 15 TB of data every night, it is crucial to develop ef-

ficient data science algorithms that are able to identify sources of interest [Djorgovski et al., 2022]. In this work, our aims are as follows:

1. Describe the multi-wavelength, spectroscopic and variability properties of a clean BAL QSO sample.

2. Test the usefulness of multimodal learning for the task of identifying BAL QSOs in large time-domain surveys such as the LSST.

To this end, we characterize a clean BAL QSO sample by constructing its mean SED and composite spectrum, and analyzing its variability. Additionally, given the relevance of the CIV absorption features in the characterization of BAL QSOs, we also aim to investigate its effect on the mentioned properties. Additionally, we use the clean sample to build and test spectrum-assisted light-curve classifiers with a multimodal learning approach.

In Chapter 2, the studied sample is presented and analyzed. In Chapter 3, we describe the ML experiments done. Finally, we present a summary of our results and future prospects in Chapter 4.

Throughout this work, we assume a flat ΛCDM cosmology with parameters $H_0 = 70$ km s^{-1} Mpc^{-1}, $\Omega_\Lambda = 0.70$ and $\Omega_M = 0.30$.

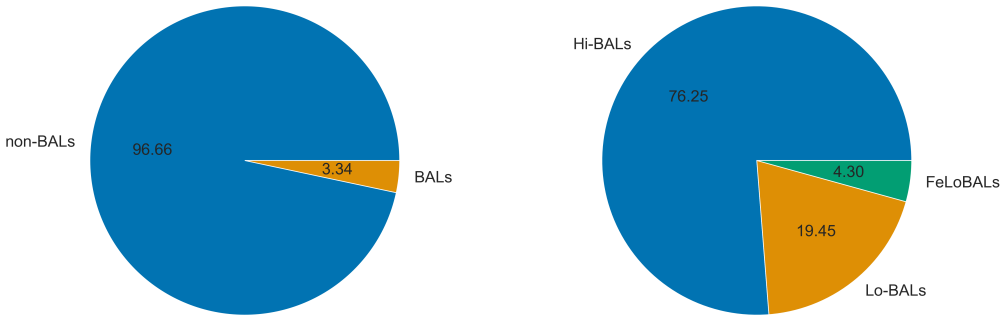
For the implementation of data science and machine learning methods [e.g. Géron, 2019, Ivezić et al., 2014], the `scikit-learn` [Buitinck et al., 2013, Pedregosa et al., 2011], `keras` [Chollet et al., 2015] and `tensorflow` [Abadi et al., 2015] libraries were used.

Chapter 2

Properties of the Reference Sample

2.1. Presentation of the Sample

The studied sample consists of 1419 BAL QSOs and 41086 non-BAL QSOs selected by Naddaf et al. [2023] from the SDSS DR7 QSO catalogue [Shen et al., 2011]. Our BAL QSOs are classified according to their ionization level: there are 1082 Hi-BALs, 276 Lo-BALs and 61 FeLoBALs. In addition, 95 non-BAL QSOs are labeled as Mini-BALs, which have a null BI and far too low outflow velocities for a BAL, but present smooth absorption features like other BALs due to blended narrow absorption lines, and thus could contaminate BAL samples [Hamann et al., 2013]. Since the spectra of the sources in our sample were visually inspected, it is as clean as possible. Figure 2.1 illustrates the proportion of BAL classes and sub-classes present in the studies sample and Figure 2.2 shows the distribution of our sources in the sky.



(a) (b)

Figure 2.1: Pie charts illustrating breakdown of sub-classes in our sample. In panel (a), there is the proportion of BALs to non-BALs in our sample, and in panel (b) the proportion of Hi-BALs to Lo-BALs and FeLoBALs.

The sources were selected with the following criteria: a median signal-to-noise ratio $S/N >$ 10 per pixel in the rest-frame $\lambda 2700 - 2900$ Å MgII region, and a measured black hole mass estimate from MgII $\lambda 2800$ Å.

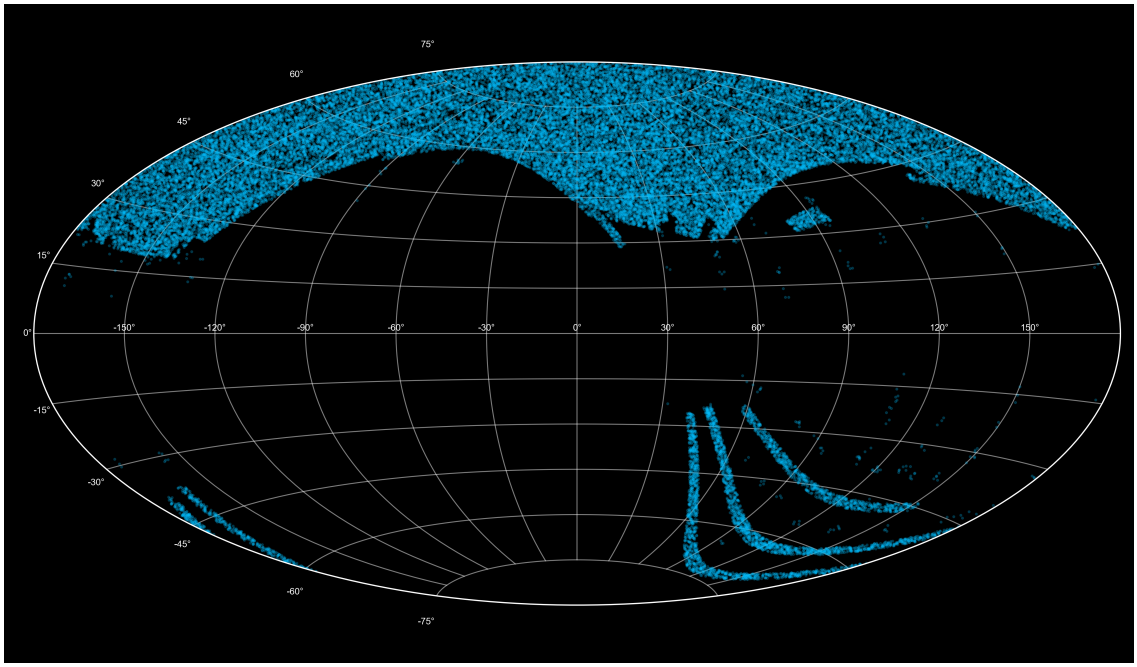
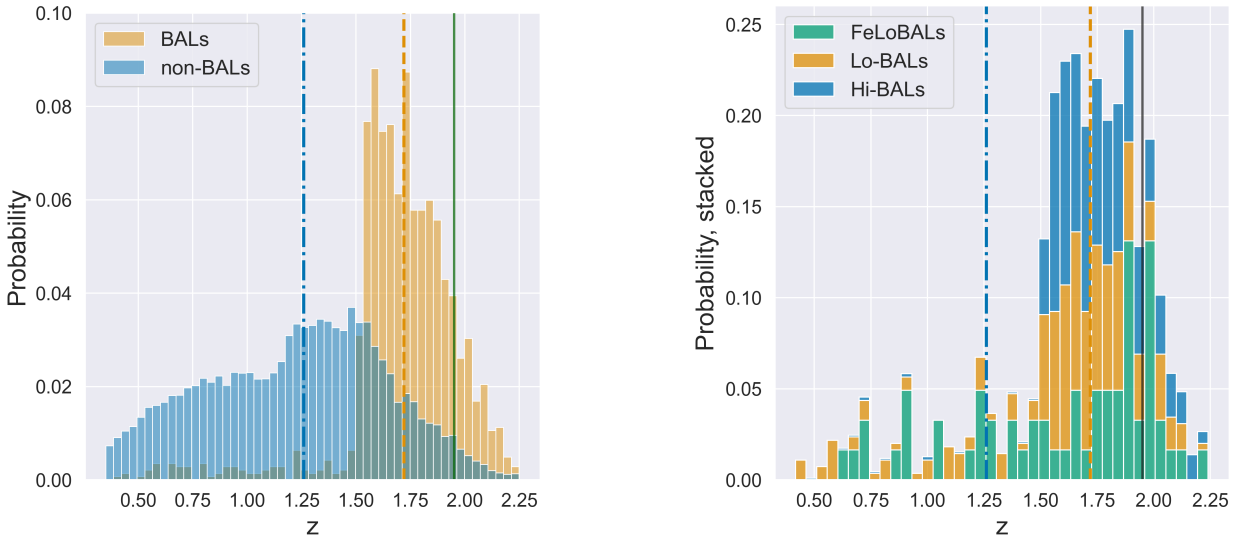


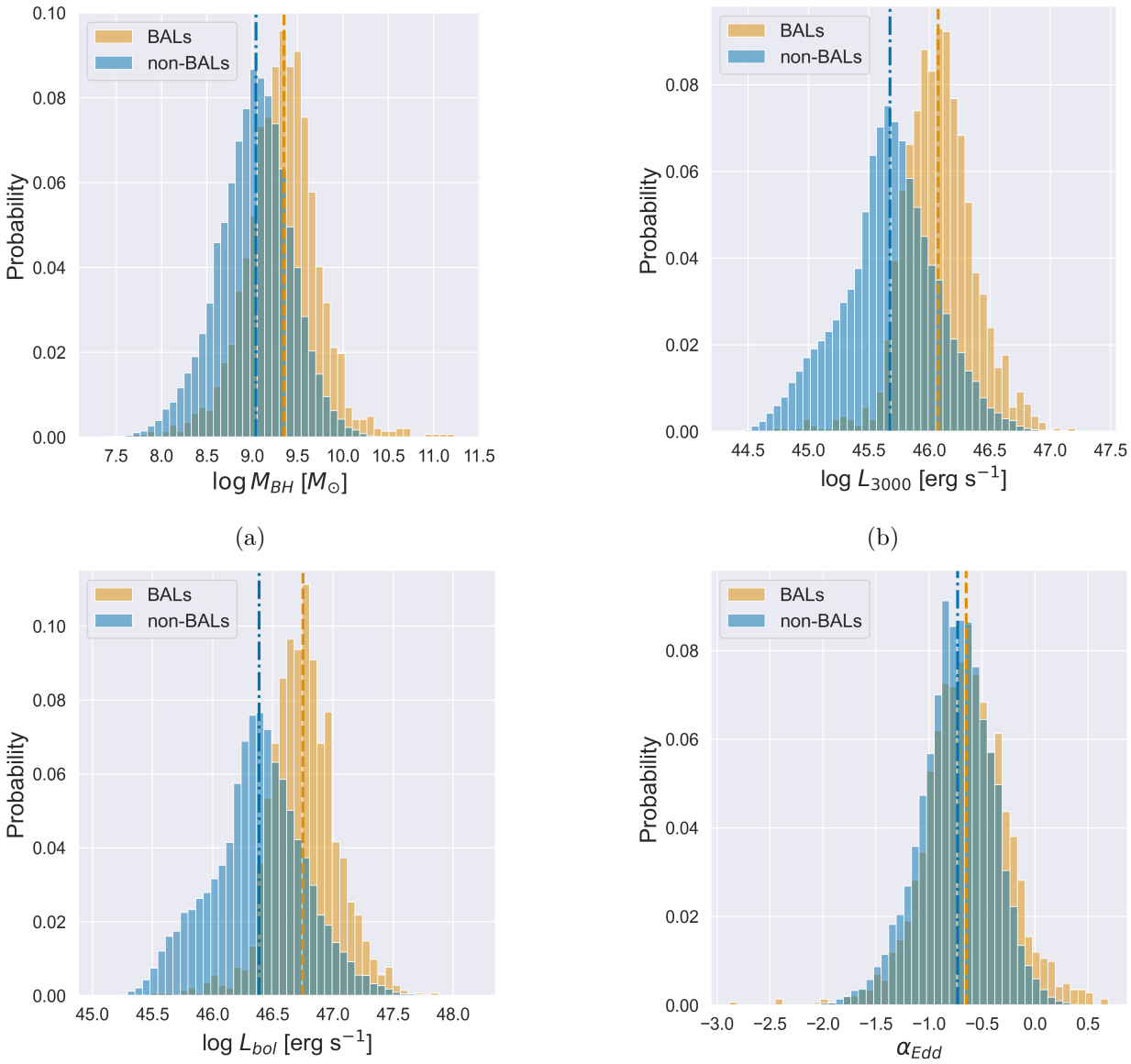
Figure 2.2: Sky map of the sample.



(a) (b)

Figure 2.3: Redshift z distribution by class in panel (a) and by BAL ionization class in panel (b). The vertical dotted and dash-dotted lines indicate the median values, and the solid black line marks the cutoff for the fully-in-g sample.

As can be seen in Figure 2.3, the vast majority of BAL QSOs lie beyond redshift $z = 1.5$, since reliable identification of Hi-BALs require the presence of the AlIII region in the spectra. The observed optical spectrum from SDSS does not cover this rest-frame wavelength region at $z < 1.5$, and so Hi-BALs are missed. In this redshift range, there are mostly Lo-BALs [see Appendix B in Naddaf et al., 2023].



(a) (b) (c) (d)

Figure 2.4: Distribution of the physical properties of our sample by class. The distribution of black hole mass estimated from Mg II is shown in panel (a), monochromatic luminosity at 3000 Åin panel (b), bolometric luminosity in panel (c), and Eddington ratio α_{Edd} in panel (d). The vertical lines indicate the median values.

Figure 2.4 shows that BAL QSOs have a systematically higher BH mass estimate, bolometric and monochromatic luminosity. Indeed, Naddaf et al. [2023], Sniegowska et al. [2023] found that BAL effects are significantly more prevalent in QSOs with $M_{BH} > 10^8 M_\odot$, which is consistent with the BH masses seen in our sample. Furthermore, Bruni et al. [2019] found a higher incidence of BALs in the WISSH sample (WISE/SDSS selected hyper-luminous (WISSH) QSOs; see Bischetti et al. [2016]). They also a fraction of Lo-BALs of around 26%, 20 times higher than the fraction found in other samples (1.3%). This is most likely explained by radiation pressure being favored in hyper-luminous QSOs, which accelerates the outflows [Gaskell et al., 2016, Giustini & Proga, 2019]. Finally, no significant difference is seen in the

Eddington ratio of BALs and non-BALs (see Figure 2.4.d) either.

In order to investigate the effect of the CIV absorption on the SEDs, composite spectra and variability of BAL QSOs, we have defined a sub-sample such that the troughs blueward from the CIV λ1548.9Å line land fully within the g-band of SDSS. This results in a redshift cutoff of $1.95 \leq z$, marked with a black vertical line in Figure 2.3. This subsample has 193 BALs and 1167 non-BALs, and will be called the "fully-in-g" BAL sample throughout this chapter. The rest of the BAL QSOs (i.e. those which CIV absorption troughs land outside of the g band), will be called the "not-in-g" BAL sample.

2.2. Spectral Energy Distributions

Here, we aim to search for any peculiar difference between the SED of the BAL QSOs and the non-BAL QSOs in our sample. We do this by constructing the mean SED of these sub-samples.

2.2.1. Multi-Wavelength Data

By construction, all sources in our sample have *ugriz* magnitudes available in SDSS. We gathered mid-IR photometry measurements from the Wide-field Infrared Survey Explorer (WISE; Wright et al. [2010]) in the bands at $\lambda = 3.4, 4.6, 12$ and $22\mu m$. All four bands were available for 98.94% of BALs (1404 objects), and for 98.88% of non-BALs (40624 objects) in our sample. We also recovered measured magnitudes in the near-IR from the Two-Micron All Sky Survey (2MASS; Skrutskie et al. [2006]) and the UKIRT Infrared Deep Sky Survey (UKIDSS; Lawrence et al. [2007]), through cross-matching with the catalogs built by Krawczyk et al. [2013] and Lyke et al. [2020]. When measurements from both surveys are present, UKIDDS is preferred since it is deeper than 2MASS (e.g. in the K band, 2MASS has a limit of 15.50 magnitudes, and for the UKIDDS Deep Extragalactic Survey this value is 20.8). Finally, 51.02% of BALs in our sample (724 objects) have available J, H and K band magnitudes from either 2MASS or UKIDDS, and 25.86% (367 objects) have a Y band magnitude available from UKIDDS. In sum, 295 BALs and 7777 non-BALs have available magnitudes in all 13 bands.

2.2.2. Methods

To build our SED, we follow the same procedure as Saccheo et al. [2023] and Krawczyk et al. [2013], using the code written by the former available on GitHub[2].

2.2.2.1. Corrections

The first step is to account and correct for effects that are external to the emission from the QSO and may modify the result. In particular, we address the absorption by the intergalactic medium (IGM), the contributions of emission lines from the BLR and NLR, and from the host galaxy.

Lyman α Absorbers in the Intergalactic Medium

Neutral hydrogen clouds in the IGM along the line of sight cause significant absorption toward shorter wavelengths than the Lyα 1216 Å line [Lynds, 1971]. With a given optical

[2] https://github.com/ivanosaccheo/my_functions

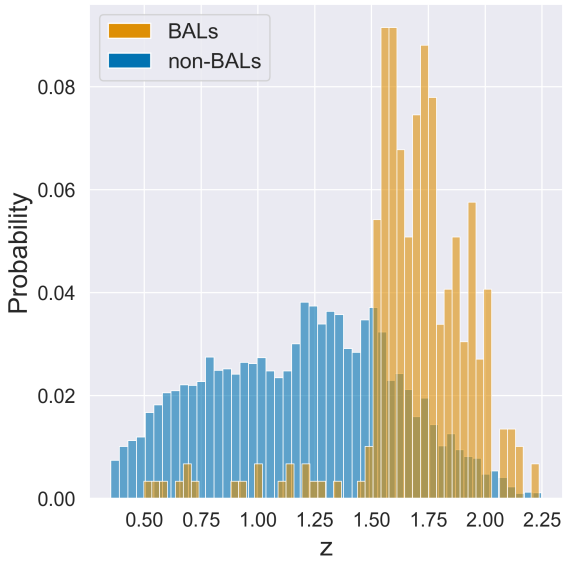


Figure 2.5: Redshift z distribution for the QSOs used for building our mean SEDs.

depth $\tau(z, \lambda)$, the magnitude offset can be derived by convolving it with the transmission curve from each filter S_λ and the continuum flux F_λ as follows:

$$\Delta m_{IGM} = -2.5 \log \frac{\int \lambda F_\lambda e^{-\tau(z,\lambda) S_\lambda d\lambda}}{\int \lambda F_\lambda S_\lambda d\lambda} \tag{2.1}$$

To this end, the IGM model by Inoue et al. [2014] was used to obtain an estimate of the optical depth τ. They describe the absorption to be due to two separate components: the Lyα forest, which dominates at column densities of $\log(N_{\rm HI}/{\rm cm}^{-2}) < 17.2$, and the damped Ly$\alpha$ systems, dominating at $\log(N_{\rm HI}/{\rm cm}^{-2}) \geq 20.3$. To correct only for the minimum needed, and also because we cannot tell whether they are present in our photometry, we assume there is no contribution from the damped Lyα component. Furthermore, to model the continuum in the UV-to-optical, a single power-law of $F_\lambda \propto \lambda^{-1.56}$ is used [Vanden Berk et al., 2001]. This calculation is done for all five of the SDSS filters.

Emission Lines

The presence of strong emission lines can overestimate the measured apparent magnitude if, at a certain redshift, they fall onto a given filter. In order to correct for this effect, we use a mock continuum flux $F_\lambda(c)$, and add the contribution of the 13 strongest lines as measured by Vanden Berk et al. [2001]. Then, the magnitude offset can be calculated as:

$$\Delta m_{EL} = -2.5 \log \frac{\int \lambda F_\lambda (c\&l) S_\lambda d\lambda}{\int \lambda F_\lambda (c) S_\lambda d\lambda}, \tag{2.2}$$

where S_λ is the transmission curve of a given filter. The continuum flux $F_\lambda(c)$ is modeled with a single power-law as mentioned above. The emission lines are described as a Gaussian profile with equivalent widths and FWHM obtained from Vanden Berk et al. [2001]. Note that this correction does not account for the skewness of the line profiles nor the dependence on equivalent width. We applied it to all 13 bands in our data.

Host Galaxy

Finally, the contribution from the host galaxy must be subtracted from the total luminosity $L_{\rm tot} = L_{\rm AGN} + L_{\rm host}$. For QSOs with $\log L_{5100} < 44.75$, the relation found by Richards et al. [2006] can be used (see also Berk et al. [2006]):

$$\log L_{6165,\rm host} = 0.87 \log L_{6165,\rm AGN} + 2.887 - \log \frac{L_{\rm bol}}{L_{\rm Edd}}, \tag{2.3}$$

where $L_{\rm bol}/L_{\rm Edd}$ is taken to be equal to 1 since it accounts for the minimum correction required. For QSOs with $\log L_{5100} < 45.053$, the relation for high-luminosity QSOs found by Shen et al. [2011] should be used instead:

$$\frac{\log L_{5100,\rm host}}{\log L_{5100,\rm AGN}} = 0.8052 - 1.5502x + 0.9121x^2 - 0.1577x^3, \tag{2.4}$$

where $x + 44 \equiv \log L_{5100,\rm tot}/[\rm erg\ s^{-1}]$.

The monochromatic luminosity at 5100Å was recovered from the data by interpolating between the two closest bands available.

2.2.2.2. Obtaining the Mean SED

After applying the corrections, the data is interpolated over a grid of wavelengths with $\Delta(\log \lambda) = 0.02$ from 912 Å and 15.8 μm to obtain a homogeneously binned SED at luminosities L_i. The uncertainties are derived by interpolating the upper bounds of the luminosity values, i.e. $L_i + \sigma_i$, over the wavelength grid, and then subtracting the main luminosities.

Then, the mean SED is computed at each wavelength bin with the weighted geometric mean:

$$\overline{\lambda L} = \exp\left(\frac{\sum_i^N \log(\lambda L_i) w_i}{\sum_i^N w_i}\right), \tag{2.5}$$

where $w_i \equiv (\lambda L_i/\sigma_i)^2$ are the weights and N is the number of data points. The uncertainties in the SED are given by the geometric variance:

$$\sigma^2 = \frac{\sum_i^N \log\left(\frac{\lambda L_i}{\overline{\lambda L}}\right)^2}{N-1} \tag{2.6}$$

2.2.2.3. Treatment of Errors

We found that WISE uncertainties reported by Lyke et al. [2020] are far too low to be reliable. Reaching errors as low as 6×10^{-4} magnitudes and 0.005 % of uncertainty for the W1 band, these values are well below the WISE systematic uncertainty of ±1.5 % [Wright et al., 2010], which causes problems in our SED fitting. We resolve this by restricting the distribution of errors σ_i such that values below 0.05 magnitudes are scaled up to $\sqrt{\sigma_i{}^2 + 0.05^2}$ to combine the reported errors with an additional systematic uncertainty of 0.05 magnitudes.

2.2.3. Results

In Figure 2.6, we show our mean SED derived for all BALs. The objects with redshifts below 1.5 (see Figure 2.5) are systematically dimmer, which is expected.

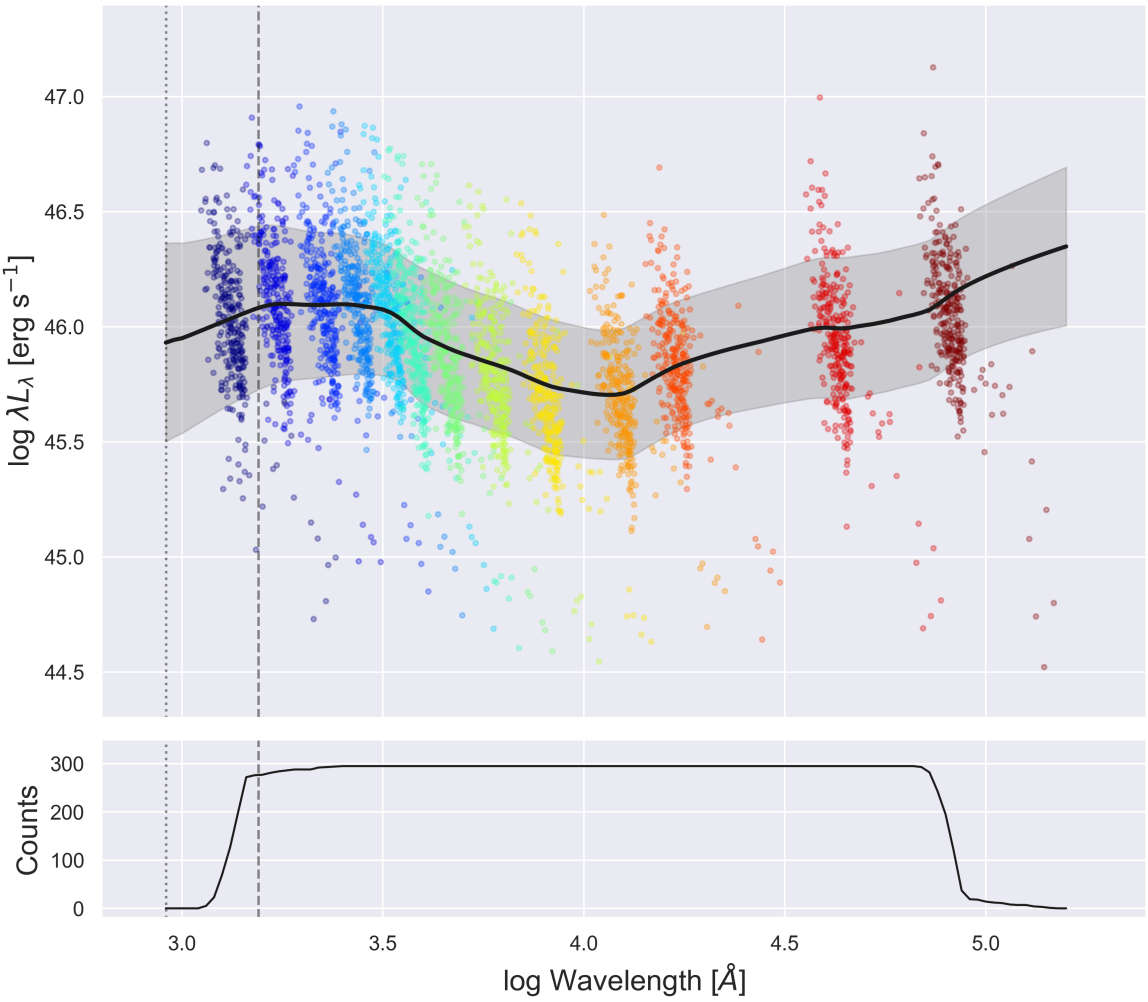


Figure 2.6: Mean SED for our BAL sample. The vertical dotted and dashed lines mark the Lyα limit at 912 Å and the CIV line at 1548.9 Å respectively. The shaded area corresponds to the 68% confidence interval.

2.2.3.1. SED for BAL QSOs vs. non-BAL QSOs

In order to compare our mean SED with the one derived from non-BALs and templates in the literature, we normalize at $\log 4.5$Å. In other words, we assume equal monochromatic luminosity at this wavelength allowing for relative comparison between the SEDs. Figure 2.7 displays our mean BAL SED compared with the one for non-BALs, as well as the template for a general QSO population by Krawczyk et al. [2013] and for high-luminosity BAL QSOs derived by Saccheo et al. [2023].

Here, we recover the redder UV-optical continuum between 1000Å and 1 μm, a behavior consistently found in the literature [e.g. Gallagher et al., 2007, Krawczyk et al., 2015, Reichard et al., 2003, Saccheo et al., 2023, Trump et al., 2006]. This is explained by higher dust extinction associated to the outflows in BAL QSOs.

At longer wavelengths, in the IR range ($\lambda > 1$ μm), we find a similar behavior to that obtained by Saccheo et al. [2023]. The BAL SED is steeper, once again pointing to a large dust contribution. It has been proposed that the outflowing gas has a dust component. By

deriving the attenuation curve for BAL dust, Gaskell et al. [2016] conclude that there is no intrinsic difference between the SED of BALs and non-BAL QSOs, and the reddening is purely due to dust. Additionally, they argue that this extra dust in BAL QSOs is associated with their high-velocity outflows, since radiation pressure on the dust will drive the outflow if the gas and dust are coupled. Zhang et al. [2014] also propose that a dust component within the outflow explains the found correlations between the NIR slope and BAL parameters. The results seen here are consistent with this explanation.

Moreover, the SED obtained by Saccheo et al. [2023] for high luminosity BAL QSOs shows a significantly flatter UV continuum than the BAL SED obtained here. This indicates that the BALs in the WISSH sample could have stronger and/or deeper absorption features. One of the QSOs in the fully-in-g sample also happens to be in the WISSH sample (WISSH01, SDSS 004527.68+143816.1). Additionally, 18 out of the 38 BAL QSOs used to compute the fully-in-g BAL SED (see Figure 2.8) have a high bolometric luminosity ($\log L_{bol}/[\mathrm{erg\ s^{-1}}] \geq 47.0$). Therefore, it is not possible to say that the difference seen in the UV continuum is due to a behavior characteristic to high luminosity BAL QSOs only, and it is most likely due to a selection effect.

We also note that the mean SED of non-BALs is similar in shape to that of Krawczyk et al. [2013], plotted with a dashed-dotted line in the plot. The only notable difference is a steeper decline between ~ 4000Å and 1 μm and a more pronounced dip at $\lambda \sim 1.3\mu$m. This difference is most likely due to selection effects, such as differences in the redshift distributions of both samples, or the sample size (their template was computed with 119652 QSOs, whilst our non-BAL SED with 7777 sources).

2.2.3.2. SED for BALs in the Fully-in-g and Not-in-g Samples

Figure 2.8 shows the SED obtained for the BALs in the fully-in-g BAL sample compared to the ones in the not-in-g sample. Respectively, 38 and 257 BAL QSOs with detections in all 13 bands were used for the computation of each of these SEDs. Given the narrow redshift range and low number of objects, their SEDs are very smooth, with nearly straight regions.

When normalizing and assuming equal monochromatic luminosities at $\log 4.5$Å, we see the obtained SEDs for both sub-samples are generally alike in shape, with the fully-in-g BAL SED being slightly dimmer. The main difference between them is that the fully-in-g BAL SED has a more pronounced dip at ~ 1 μm, which can indicate a larger relative difference in temperature between the accretion disk (dominant in the UV and optical) and the dust of the torus and the outflows (dominant in the NIR). Between the dip and the normalization point, the fully-in-g BAL SED is minimally steeper, and after it it is brighter. The IR W4 band lands at $6.5\lambda < 10.4$ μm for both samples given a median redshift of 2.01 for the fully-in-g and of 1.68 for the not-in-g samples. Thus, this brighter portion in the fully-in-g SED could possibly indicate stronger dust extinction which is not necessarily fully explained by the redshift selection of the sub-samples alone. However, given the limited number of objects used to compute the fully-in-g BAL SED, it is a possibility that these sources are by chance at an inclination angle such that the outflow is covering a larger fraction of the QSO. Further studies are needed to fully explain the origin of this difference in IR emission.

2.3. Spectra

The characterization of the BAL QSOs in our sample would not be complete without studying their spectroscopic properties. In this Section, we describe their spectral parameters,

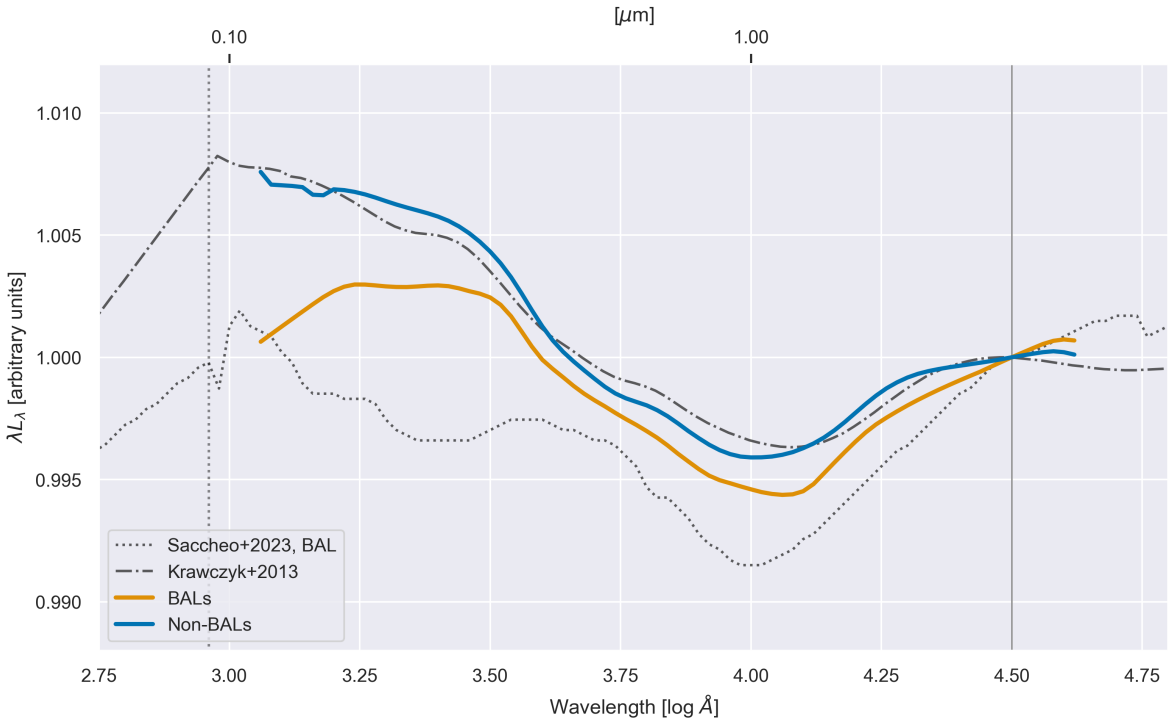


Figure 2.7: Mean SED computed for our BAL (in orange) and non-BAL (in blue) samples compared to the ones computed by Krawczyk et al. [2013] for a general QSO population (dash-dotted line), and by Saccheo et al. [2023] for high-luminosity BALs (dotted line). They are normalized to $\log 4.5$Å, marked by a solid vertical line.

and build composite spectra to ease the analysis.

We fetch SDSS spectra from the 18th Data Release (SDSS DR18; Almeida et al. [2023]) through the SDSS Science Archive Server (SAS)[3]. They were recovered for all objects except for 29 non-BALs, one Hi-BAL, and one Lo-BAL. Figure 2.9 displays an example from our sample spectrum for each of the BAL ionization classes.

Furthermore, Figures 2.10 display the distribution of the C IV emission equivalent width and FWHM for BALs and non-BALs, and 2.11 the ones of BI and AI for BALs only, as measured by Lyke et al. [2020]. We note that the CIV parameters are similar between BALs and non-BAL QSOs, indicating that the key difference between them is not CIV emission, but the troughs blueward of the CIV rest-frame wavelength. Figure 2.11 shows that the AI and BI distributions are skewed to the left, indicating BALs with more extreme absorption are less common. Additionally, out of the 7909 non-BAL QSOs with available AI and BI measurements, 116 have positive BI values, with a minimum of 2.00, a median of 187.69 and a maximum of 109060.02 km s^{-1}) . These non-BALs were flagged and excluded from the ML experiments described in Chapter 3.

2.3.1. Composite Spectra

The composite spectrum of a large sample is an intuitive way of visualizing its characteristic continuum and absorption. In this section, we build composite spectra to compare BAL

[3] https://data.sdss.org/sas/dr18/spectro/sdss/redux

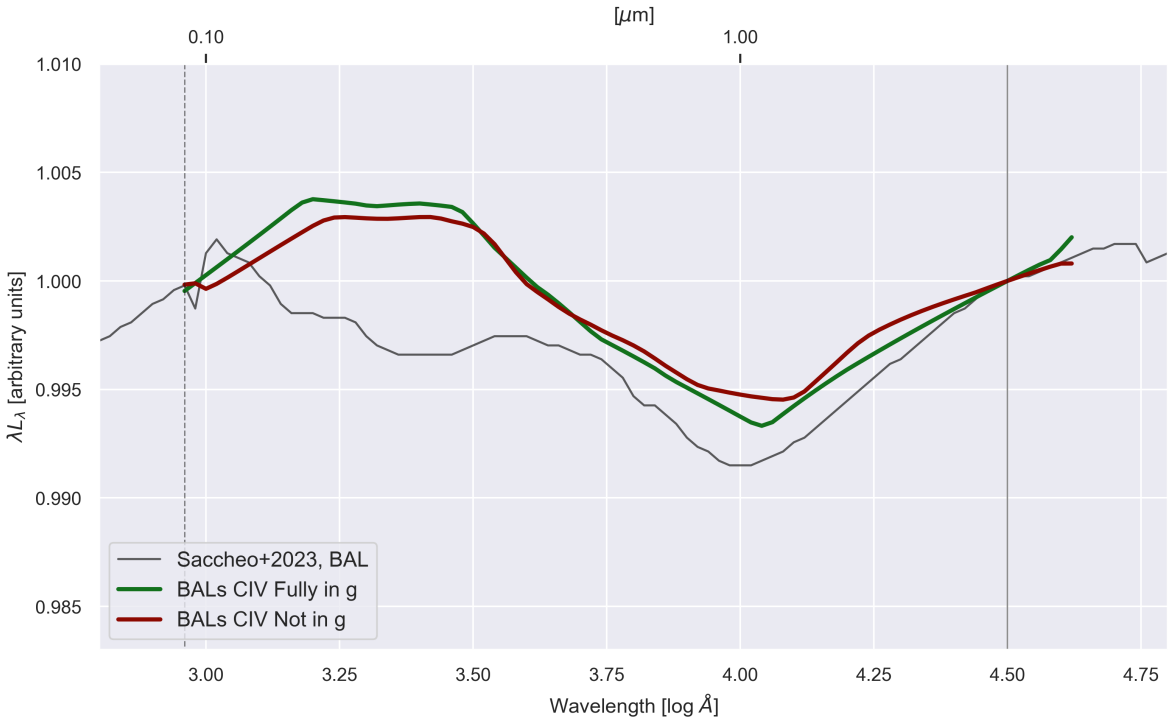


Figure 2.8: Mean SED computed for our fully-in-g and not-in-g BAL samples compared to the one computed by Saccheo et al. [2023] for high luminosity BAL QSOs. They are normalized to $\log 4.5$Å, marked by a solid vertical line.

sub-classes with each other, and to check for changes with BI.

2.3.1.1. Methods

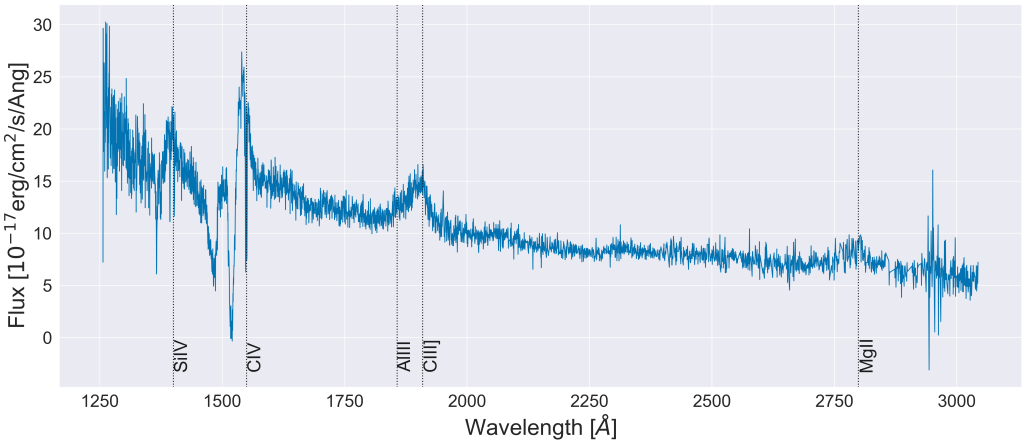
With some minor correction, we use use the code written by Hans Klaufus available on GitHub [4], a script tailored specifically for building a composite spectrum of a given selection of SDSS QSO spectra.

After reading all the given spectra, it corrects them for redshift such that they are in the rest-frame. The code was modified such that it uses the redshift value reported by the reference paper of our sample [Naddaf et al., 2023] instead of that found within the spectrum files from the SAS.

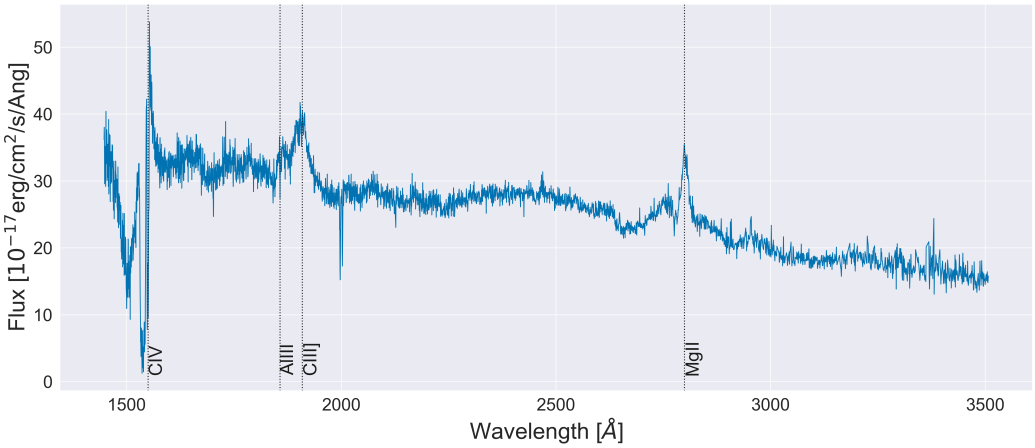
Then, it sets the flux value of bad pixels to zero, such that when the spectra are later combined, these null values will not contribute to the mean flux in the given wavelength bin, effectively excluding them from the resulting composite spectrum. Bad pixels are identified with the `and_mask` reported by SDSS[5]. Note that for composite spectra created for a small number of objects, this could lead to gaps if the bad pixels are too prevalent.

Next, the spectra are normalized. In increasing order of redshift, pairs of consecutive individual spectra are taken and the mean flux of each of them is calculated. A normalization factor equal to the ratio of these mean fluxes is used to normalize this pair of spectra. This continues until the fluxes and inverse variances of all of the individual spectra have been

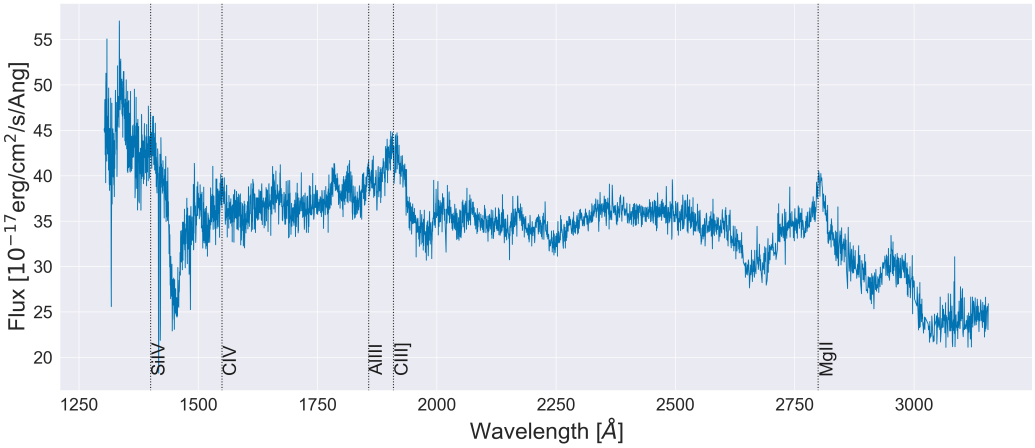
[4] https://github.com/hklaufus/CompositeSpectrum
[5] See https://www.sdss4.org/dr17/spectro/quality/#SpectrumQuality



(a) Hi-BAL Spectrum: SDSS 100021.72+035116.5, at $z = 2.021$.

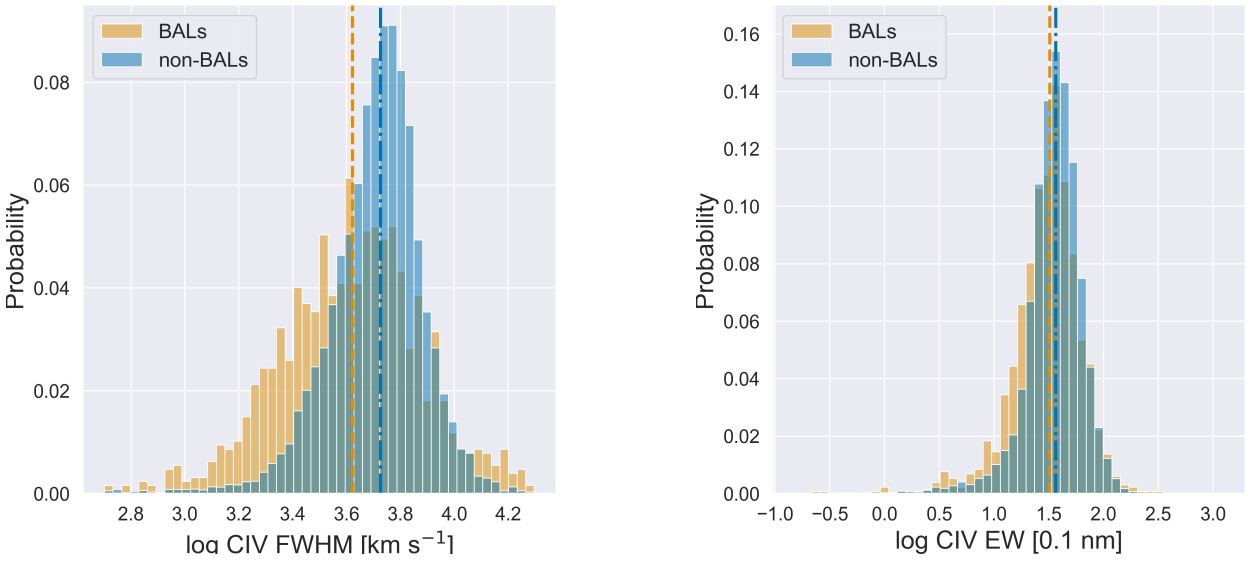


(b) Lo-BAL Spectrum: SDSS 121440.27+142859.1, at $z = 1.6245$.



(c) FeLo-BAL Spectrum: SDSS 143752.75+042854.5, at $z = 1.9188$.

Figure 2.9: Examples of individual spectra by BAL ionization class.



(a) (b)

Figure 2.10: Properties of the CIV line profile as measured by Shen et al. [2011]. Panel (a) shows the FWHM of the CIV profile [km s^{-1}] of the available data for 1271 BAL and 11183 non-BAL QSOs. Panel (b) shows the restframe equivalent width of the whole CIV profile [0.1 nm] of the available data for 1307 BAL and 11244 non-BAL QSOs.

normalized relative to each other in this way. This ensures that all spectra are consistently in the same relative flux range, even if some of them do not overlap with each other over the rest-frame wavelength.

Finally, it re-bins the spectra to the inputted wavelength range and grid with $\Delta(\log \lambda) =$ 1Å. We apply a geometric mean to calculate the flux, uncertainty and noise in each bin, using all available spectra in the given bin.

2.3.1.2. Results

Figure 2.12 displays the obtained composite spectrum of Hi-BALs, Lo-BALs and FeLoBALs. The characteristic shape of each class is recovered. They all show strong absorption blueward of the CIV line.

The LoBAL composite is the reddest one, with the flattest continuum, which has been previously found Gibson et al. [2009], Reichard et al. [2003]. However, in spite of being the defining feature of this class and being present in Figure 2.9.b, the absorption blueward of the MgII λ2799 line is only slightly noticeable. This is most likely because of the varying depths, blue-shifts and widths of the absorption that wash out the individual absorption features when computing the mean.

Moreover, the FeLo-BALs are quite distinct from both Hi-BALs and Lo-BALs. The FeII line absorption in this ionization class is appreciated at several wavelengths, specially at 1608.45 Å, whilst not being present for the other classes. Additionally, the close-up to the SiIV and CIV lines in Figure 2.12.b reveals that its CIV emission is more blue-shifted than for the other sub-classes. However, this could be intrinsic to the present sample here, and not necessarily to the FeLo-BAL class in general. Its continuum is also steeper, less uniform, and has excess emission in several regions, such as between the FeII lines at 2382.765 and

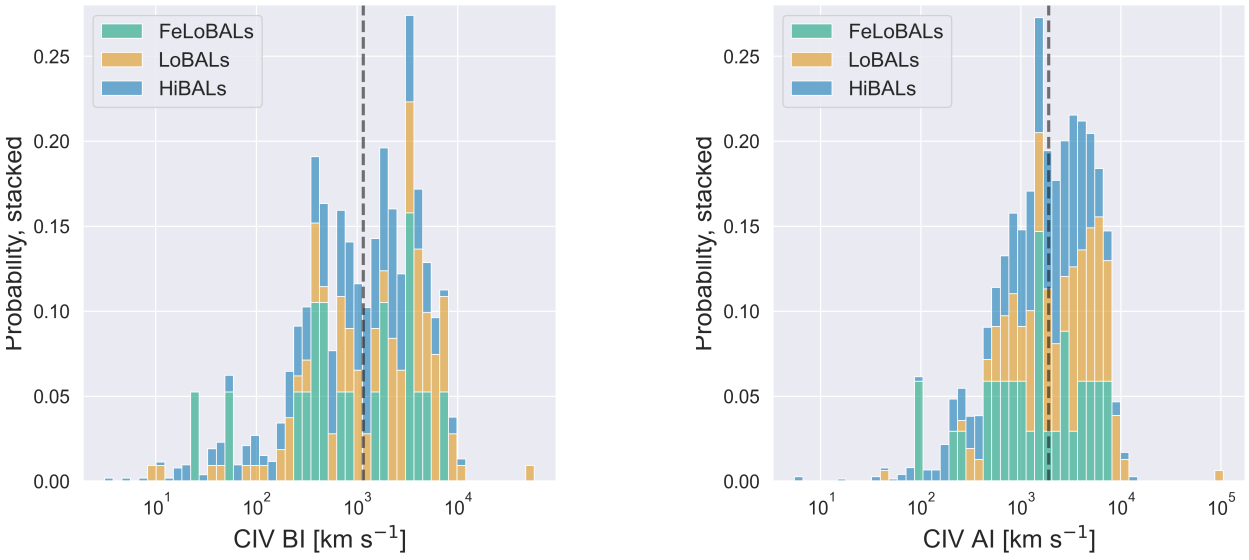


Figure 2.11: CIV BI and AI as measured by Lyke et al. [2020] in panels (a) and (b) respectively. The vertical dashed black lines are at the median values. These histograms were built with the available data for 993 BAL QSOs.

2586.650 Å, and between FeIIλ1608.5Å and AlII λ1857Å. FeLo-BALs are a rare and special class of BALs Leighly et al. [2024], Menou et al. [2001], Trump et al. [2006], which can be appreciated in our composite.

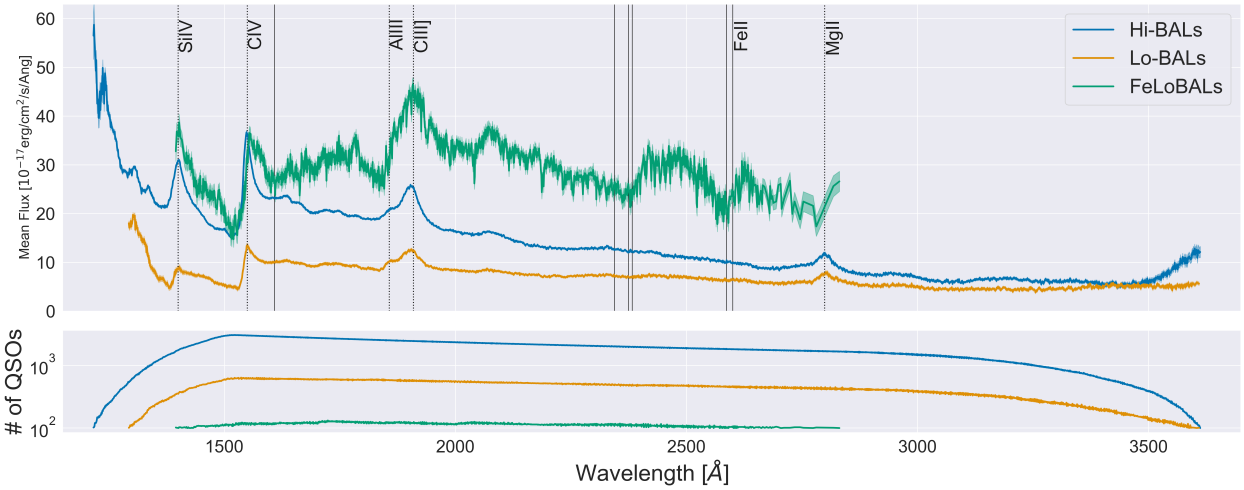
Furthermore, Figure 2.13 displays the obtained composites for our BALs separated by BI bins corresponding to the 25th percentile, median, 75th percentile and maximum BI values of our sample. Except for the two composites with highest BI where there is no major difference in the continuum, we see that the higher the BI of the sub-sample, its composite has a flatter continuum. This is consistent with the idea that BALs with higher BI have stronger outflows, and thus, a higher dust extinction.

In the lower panel (Figure 2.13.b), we see that the BI does not relate directly to the depth, width or blue-shift of the absorption troughs. Indeed, the BI is a proxy for any absorption, not its properties. To fully describe the CIV absorption, one should include the number, width and blueshift of the troughs. In the interactive plots developed by Rankine et al. [2020] [6], it is intuitive to see the vast diversity of absorption and emission shapes in both BAL and non-BAL QSOs. They build a CIV emission space on its EW and blue-shift. A similar work with a detailed description of the CIV troughs would bring valuable insight into the detailed structure of the outflows in BAL QSOs, including its shape, density and strength. This could proof crucial to feedback studies in BALs.

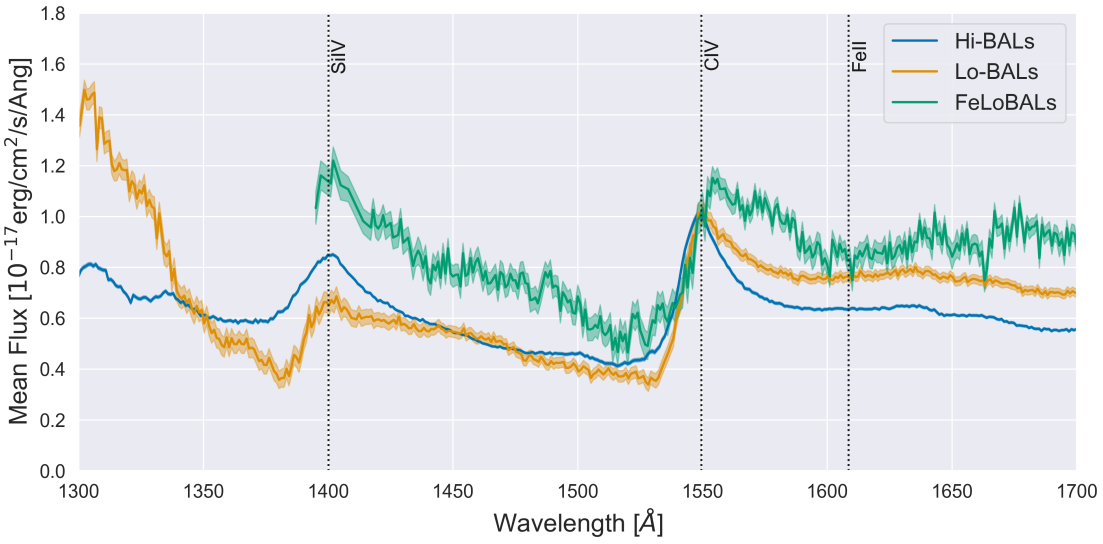
2.4. Light-Curves

In this section, we compute statistical tests to compare the features of BALs and non-BALs to confirm whether BAL QSOs indeed do not have a distinct variability behavior. We

[6] See supplementary data in https://academic.oup.com/mnras/article/492/3/4553/5707433

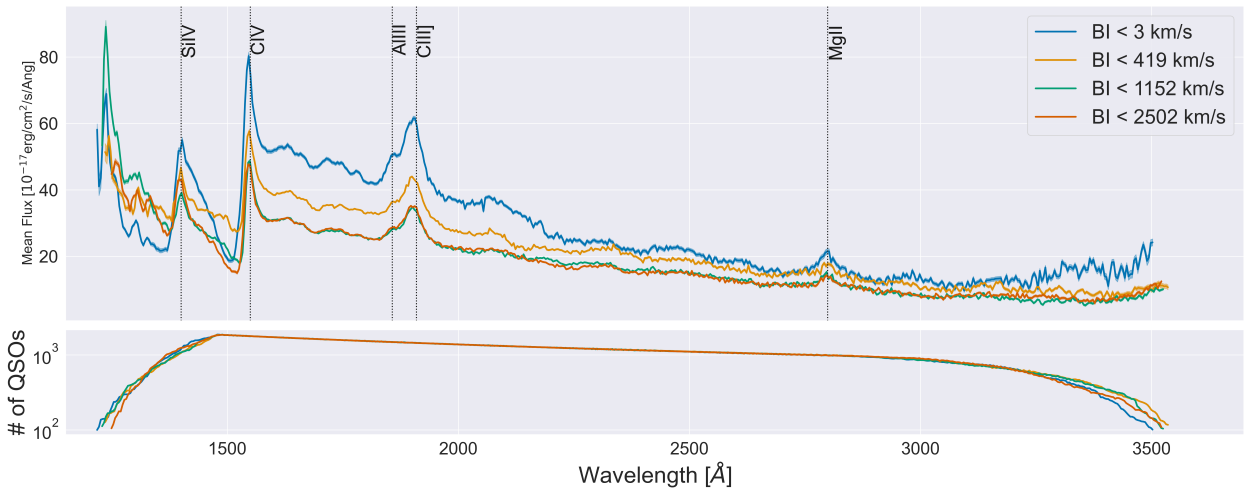


(a)

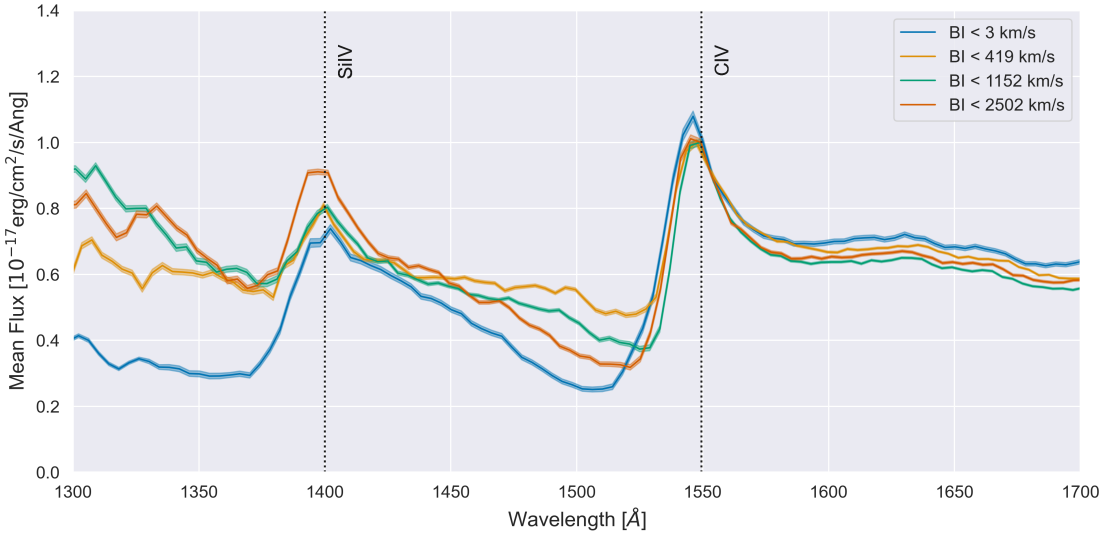


(b)

Figure 2.12: Composite spectrum of all Hi-BALs, Lo-BALs and FeLo-BALs. Panel (a): full composite spectra for each BAL ionization class at the top and the number of objects used at each wavelength bin at the bottom; only bins computed with 100 or more spectra are plotted; all the solid lines correspond to FeII lines. Panel (b): zoom in to the blue-shifted absorption troughs of the CIV and SiIV lines of the composites; they are normalized at the CIVλ1548.9 restframe wavelength.



(a)



(b)

Figure 2.13: Composite spectrum for BALs by BI bins of 419, 1152, 2502 and 57500 km s^{-1}, corresponding to the 25th percentile, median, 75th percentile and maximum of the BI distribution of our BAL sample. Panel (a): full composite spectra for each BAL BI bin at the top and the number of objects used at each wavelength bin at the bottom; only bins computed with 100 or more spectra are plotted. Panel (b): zoom in to the blue-shifted absorption troughs of the CIV and SiIV lines of the composites; they are normalized at the CIVλ1548.9 restframe wavelength.

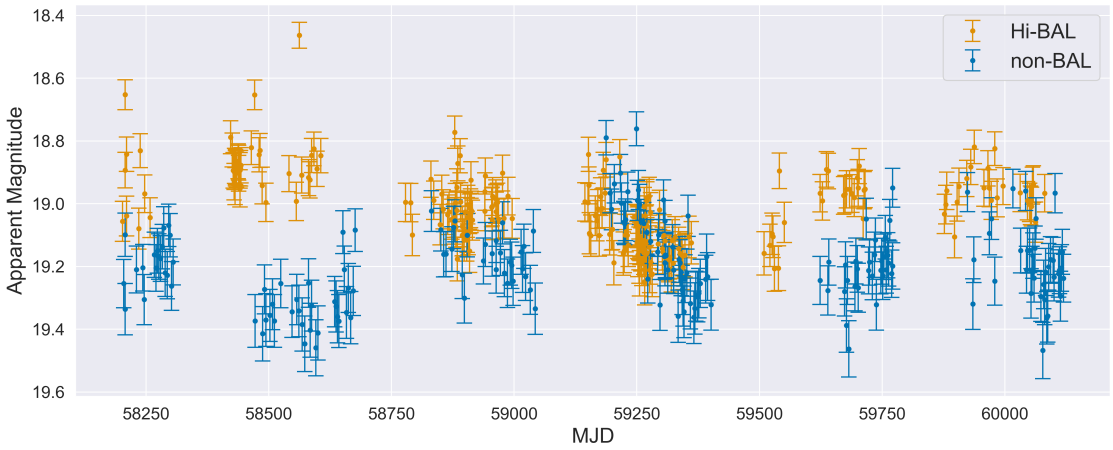


Figure 2.14: Example light-curves of a Hi-BAL QSO (SDSS 100021.72+035116.5) and a non-BAL QSO (SDSS 130650.08+002753.8).

also compare the features of BALs in the fully-in-g sample with the rest of BALs to see if it is possible to use the position of the CIV blue-shifted troughs in a given photometry filter as a map for finding BAL QSOs via variability.

2.4.1. Methods

2.4.1.1. Data

To analyze the variability of our BAL QSO sample, we have recovered their light-curves with at least four data points from ZTF DR20[7], covering observations from March 2018 to October 2023 [Masci et al., 2018]. Note that we do not use data from broker alerts and the DR data instead. This allows us to analyze the general variability behavior in BAL QSOs in spite of their obscuration, which can lead to seemingly less variability that is not always caught by alerts. Light-curves were found in the g-band for 40973 objects. Out of the 1517 objects with no light-curve, 47 were BALs. Figure 2.14 shows the light-curves of a randomly chosen Hi-BAL and non-BAL QSOs as examples.

To clean the light-curves, we excluded bad or unusable magnitudes marked with a `catflag` equal to 32768, and those with errors larger than one magnitude.

2.4.1.2. Feature Extraction

Their time-domain features were extracted using the code by [8] Malanchev et al. [2021], used by the ZTF and future LSST brokers ANTARES [Matheson et al., 2021], AMPEL [Nordin, J. et al., 2019] and FINK [Möller et al., 2020]. The computed features are described as follows:

1. `Amplitude` $\star$: half the difference between the maximum and the minimum magnitude
2. `AndersonDarlingNormal` $\star$: test of whether a sample of data was drawn from a given probability distribution

[7] See https://irsa.ipac.caltech.edu/data/ZTF/docs/releases/dr20/ztf_release_notes_dr20.pdf
[8] See https://github.com/light-curve/light-curve-python

3. `BeyondNStd` $\star$: fraction of observations beyond $n\sigma$ from the mean magnitude; n was set to 1.

4. `Cusum` $\star$: series of cumulative sums.

5. `Duration`: time-series duration.

6. `Etae` $\star$: adapted Von Neummann η for unevenly sampled time-series; ratio of the mean of the squares of successive magnitude differences to the variance of the time-series.

7. `ExcessVariance` $\star$: measure of the intrinsic variability amplitude.

8. `InterPercentileRange` $\star$: difference between two specified percentiles in a the magnitude sample of the time-series.

9. `Kurtosis` $\star$: excess kurtosis of magnitude.

10. `LinearFit` $\star$: slope, error and reduced χ^2 of a linear fit to the time-series with respect to the observation errors.

11. `LinearTrend` $\star$: slope, error and noise level of a linear fit to the time-series without respect to the observation errors.

12. `MagnitudePercentageRatio` $\star$: ratio between two percentile magnitude differences.

13. `MaximumSlope` $\star$: maximum slope between two consecutive observations.

14. `MaximumTimeInterval/MinimumTimeInterval`: maximum/minimum time interval between two consecutive observations.

15. `Mean`: mean magnitude.

16. `MeanVariance` $\star$: ratio between the standard deviation and the mean magnitude.

17. `Median`: median magnitude.

18. `MedianAbsoluteDeviation` $\star$: median discrepancy of the magnitudes from the median magnitude.

19. `MedianBufferRangePercentage` $\star$: fraction of observations within 10% of the median magnitude.

20. `ObservationCount`: number of data points.

21. `OtsuSplit` $\star$: difference of subset means, standard deviations and lower-to-all observation count ratio, for two subsets of magnitudes obtained by Otsu's method split, which separates data into two subsamples by minimizing intra-class variance and maximizing inter-class variance.

22. `PercentAmplitude` $\star$: largest percentage difference between a magnitude and the median.

23. `PercentDifferenceMagnitudePercentile` $\star$: ratio of a given inter-percentile range to the median magnitude.

24. `Periodogram` $\star$: peaks of Lomb–Scargle periodogram and the periodogram itself as a meta-feature.

25. `ReducedChi2` $\star$: reduced χ^2 of the magnitudes.

26. `Skew` $\star$: skewness of magnitude.

27. `StandardDeviation` $\star$: standard deviation of the magnitude.

28. `StetsonK` $\star$: Stetson K coefficient, a robust measure of the kurtosis.

29. `TimeMean`: mean time.

30. `TimeStandardDeviation`: standard deviation of time points.

31. `WeightedMean`: weighted mean of the magnitude.

After pre-processing, the recovered time-series have a mean and median of 466 and 373 data points. The shortest one has four detections (the minimum required for the appropriate calculation of the features) and the longest one has 5052 detections. They cover 1794.53 days on average, and the shortest and longest light-curves are 3.04 and 1926.13 days long respectively. They are unevenly sampled: the maximum and minimum time intervals between consecutive detections are on average 184.84 and 0.01 days respectively. Furthermore, the mean apparent g magnitude is on average 18.85 magnitudes.

2.4.1.3. Comparison Tests

We conducted a comparison of the distributions of time-domain features with two statistical tests. A p-value (i.e. the probability that a statistical summary of the data, e.g. the sample mean difference, would be equal to or more extreme than its observed value) threshold α was set to 0.01 for all tests to gauge the statistical significance of the result [Wasserstein & Lazar, 2016]. If $p_{val} \leq \alpha$, then there is a significant difference between the compared distributions. However, due to the tricky interpretation of the p-value, other criteria were used instead when possible. The statistical tests are described as follows:

1. Kolmogorov-Smirnov test (D_{KS}) [Kolmogorov-Smirnov et al., 1933]: whether two samples were drawn from the same distribution. It ranges from zero to one and consists of the maximum distance between the cumulative distribution functions. Thus more distinct distributions will result in a higher value. Features with $D_{KS} > 0.2$ were flagged as having different distributions by KS.

2. Levene's test (W) [Levene, 1961]: checks for homoskedasticity (i.e. homogeneity of variance) and provides a comparison of the standard deviations of the samples. The p-value was used to assess this test.

These tests were run on those features that can be interpreted to represent some aspect of variability. These are marked with a star $\star$ in the list of time-domain features (see Section 2.4.1.2). According to the mentioned criteria, if a feature is flagged to be different for the samples, we proceed to assist the comparison with plots. This visualization is key to avoid any biases caused by, for example, the imbalanced sizes of the compared samples or the p-value interpretation.

2.4.2. Results

2.4.2.1. Comparison of BAL and Non-BAL Features

Table 2.1 shows the results of the tests when comparing the features from BALs and non-BALs. According to the Kolmogorov-Smirnov flag, the excess variance and the reduced χ^2 are different for both populations. On the other hand, the Levene's test reveals that the majority of features have differing variances between the samples. However, this is not enough to claim that the feature for both populations have a significant difference. Thus, to analyse these results further, we create a visualization of those features flagged by the KS test to aid the comparison, including a histogram, violin plot, cumulative sum plot and quantile-quantile (Q-Q) plot.

Table 2.1: Results of the Kolmogorov-Smirnov and Levene tests when comparing the time-domain features from BALs and non-BALs. The flags indicate the tests indicate there is a significant difference between the given feature between the samples.

Feature	D_{KS}	$p_{val,KS}$	W	$p_{val,W}$	KS Flag	Lev. Flag
Amplitude	0.111	0.000	19.990	0.000		✓
AndersonDarling	0.056	0.000	7.437	0.006		✓
Beyond1Std	0.098	0.000	6.639	0.010		✓
Cusum	0.119	0.000	13.659	0.000		✓
ExcessVariance	0.232	0.000	38.327	0.000	✓	✓
Etae	0.117	0.000	0.030	0.862		
InterPercentileRange	0.181	0.000	77.587	0.000		✓
Kurtosis	0.120	0.000	5.007	0.025		
LinearFit_slope	0.047	0.006	0.019	0.890		
LinearTrend_slope	0.053	0.001	0.042	0.838		
MagnitudePercRatio	0.059	0.000	11.534	0.001		✓
MaximumSlope	0.022	0.554	0.151	0.698		
MeanVariance	0.194	0.000	65.458	0.000		✓
MedianAbsDev	0.180	0.000	75.310	0.000		✓
MedianBuffRangePerc	0.101	0.000	7.443	0.006		✓
OtsuSplit_diff	0.186	0.000	52.575	0.000		✓
OtsuSplit_lower	0.133	0.000	17.989	0.000		✓
OtsuSplit_upper	0.179	0.000	65.592	0.000		✓
PercentAmplitude	0.103	0.000	14.459	0.000		✓
PercDiffMagPerc	0.190	0.000	66.756	0.000		✓
Periodogram_peaks	0.045	0.008	24.773	0.000		✓
Reduced_Chi2	0.202	0.000	7.949	0.005	✓	✓
Skewness	0.086	0.000	0.758	0.384		
StandardDeviation	0.184	0.000	63.041	0.000		✓
StetsonK	0.108	0.000	0.066	0.798		

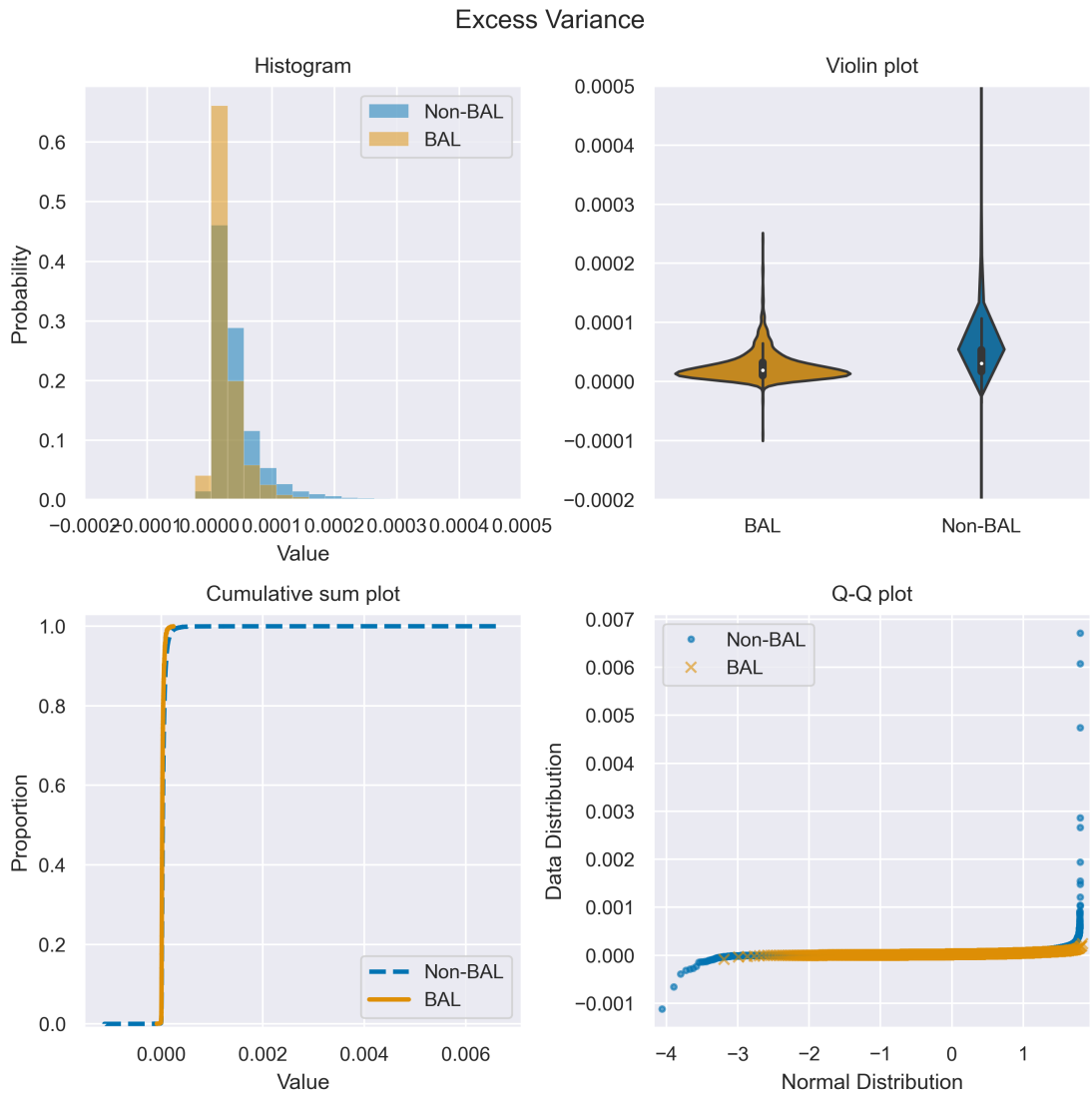


Figure 2.15: Histogram, violin plot, cumulative sum plot and Q-Q plot for the excess variance feature in BALs and non-BALs. The histogram and violin plots are zoomed into the peaks of the distribution.

Figure 2.15 shows a visualization for the excess variance feature distributions. As mentioned above, this is a measure of intrinsic variability that cannot be attributed to measurement errors or noise [Allevato et al., 2013, Sánchez et al., 2017]. In spite of being a measure of the variability amplitude able to flag intrinsically variable sources, it can depend and be biased on the structure of the time-series themselves. Additionally, the excess variance can be negative if the object is not variable and/or there are large errors. We have 19 BALs and 174 non-BALs with negative excess variance, possibly indicating that not all our studied objects have intrinsic variability. Furthermore, we see that the distribution of this feature for non-BALs has heavy tails. In the violin plot, we can see that the distributions for each population are distinct. However, this difference is minimal and both samples generally behave in a similar way in all four plots. Thus, it is not possible to conclude that the intrinsic variability is significantly different between BALs and non-BALs.

The other variability feature flagged by the KS test in Table 2.1 is the reduced χ^2. Figure

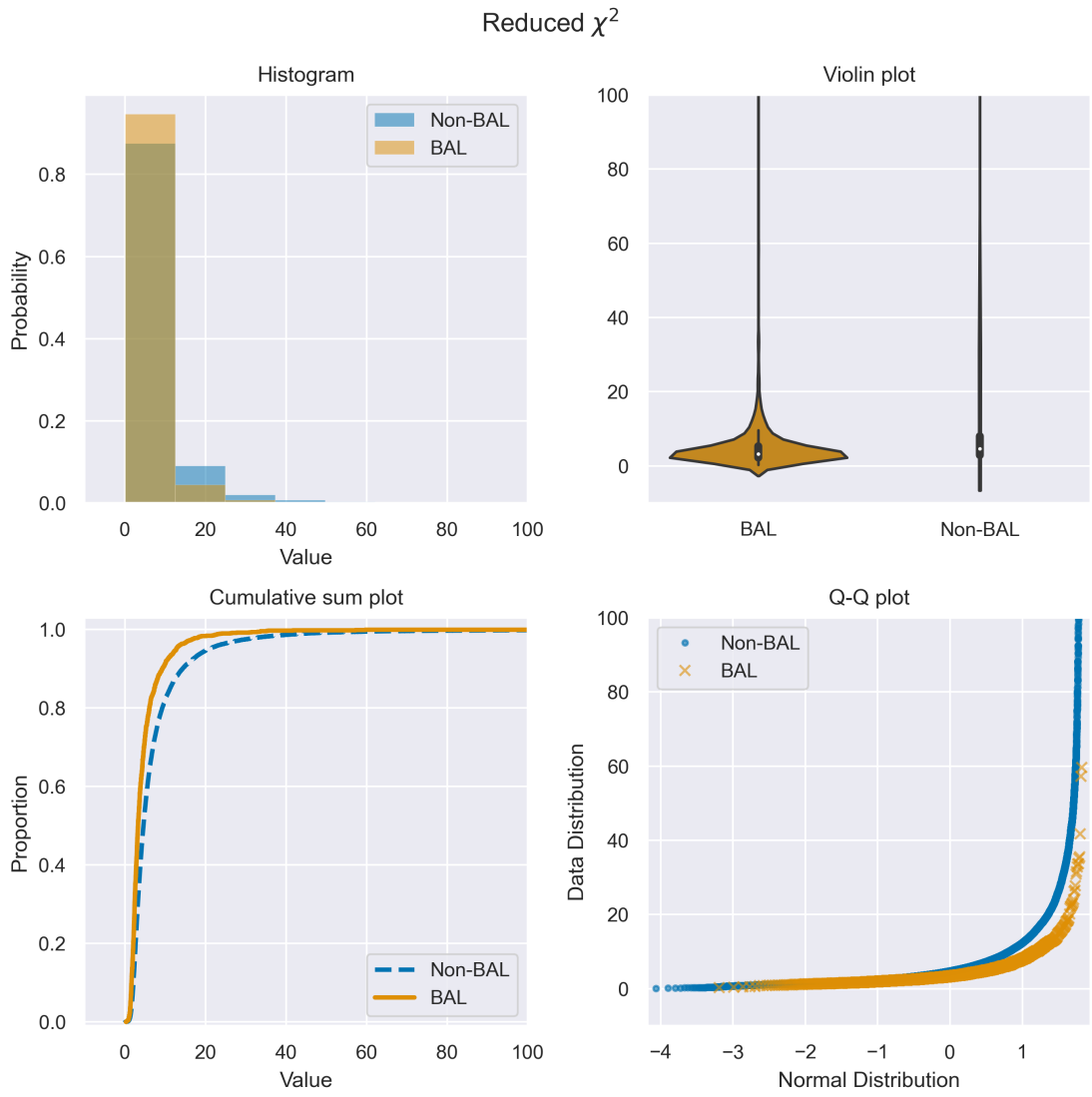


Figure 2.16: Histogram, violin plot, cumulative sum plot and Q-Q plot for the reduced χ^2 feature in BALs and non-BALs. The histogram, violin and cumulative sum plots are zoomed into the peaks of the distribution.

2.16 shows the comparison plots. This feature also is measure of variability, but unlike the excess variance, it takes into account the observation uncertainties. In the Q-Q plot, we can see that for the non-BALs, there is an extremely elongated tail that reaches two larger orders of magnitude than the median. However, the other three plots reveal there is not a significant difference between the reduced χ^2 of BALs and non-BALs.

Overall, results indicate there are no major differences between the BAL and non-BAL variability feature distributions.

2.4.2.2. Comparison of Fully-in-g BAL and non-Fully-in-g BAL Features

We also compared the features from the fully-in-g BAL sample with the rest of the BALs. In Figure 2.17, we see the transmission files for the g filter in SDSS and ZTF overlap in the majority of their wavelength ranges, and their effective wavelengths are close, differing only by 74.70Å. Therefore, it is possible to compare the fully-in-g BAL sample with the rest of

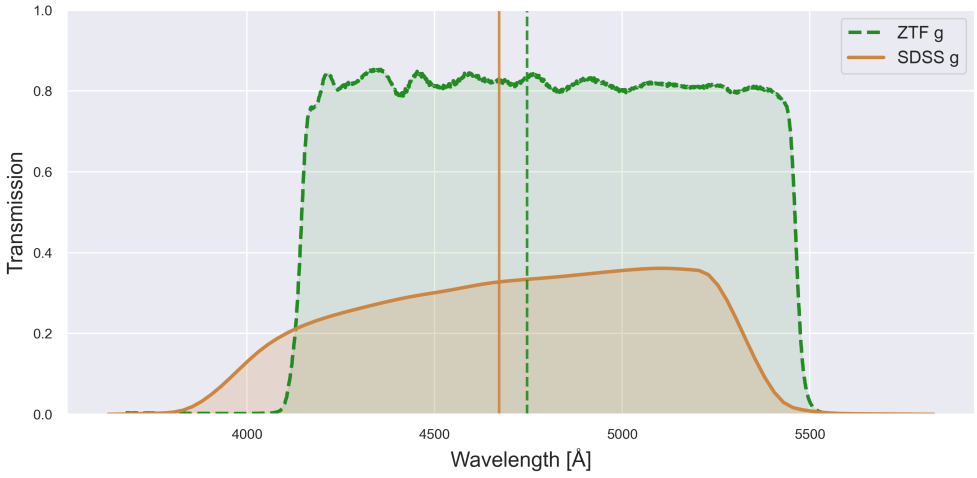


Figure 2.17: Transmission curves of the SDSS and ZTF g filters. The vertical lines are at the effective wavelengths of 4671.78Å and 4746.48Å respectively.

the BALs in a similar way as done in Sections 2.2 and 2.3.

If a significant difference is found, this could be used as a proxy for mapping the CIV into different filters in variability photometry by redshift. Thus, depending on the redshift of the studied sample, a different filter can be looked into to see if the same difference is found.

However, when comparing the features of these sub-samples of BALs, only the standard deviation of the lower subset defined by the Otsu thresholding algorithm is flagged to be different for the compared samples. This algorithm was introduced by Otsu [1979] to analyze images and find an optimal boundary between the foreground and background. In the case of astronomical light-curves, the threshold separates the baseline variability of an object from flares and periods of higher brightness. Then, the standard deviation of the lower subset can be interpreted as the standard deviation of the baseline variability. We plotted a visual comparison between this feature for BALs in the fully-in-g sample and the rest of BALs in Figure 2.18. Although the distribution for BALs not-in-g is slightly more skewed, this difference is not significant as revealed by the histogram and violin plot.

There is also no overall major difference between the feature distributions of BALs which CIV throughs land within and without the g filter. The found differences seen in Figures 2.15, 2.16 and 2.18 are minimal

2.4.3. Future Prospects

As seen here, there is no indication that BALs and non-BALs vary in an intrinsically different way. However, given the known variability of the CIV absorption troughs [De Cicco et al., 2017, Erakuman & Filiz Ak, 2017, Gibson et al., 2008, Green et al., 2023] and the glimpses of differences found here, especially when comparing features in BALs which CIV troughs land within and without the g filter, indicate there is possibly a difference to be found. This is a challenging task and requires a higher order of complexity than the one presented here.

A possibility for further investigation is to use ML, which could catch deeper level differences. A the `mTAN` [Shukla & Marlin, 2021] ML model could be used to obtain a charac-

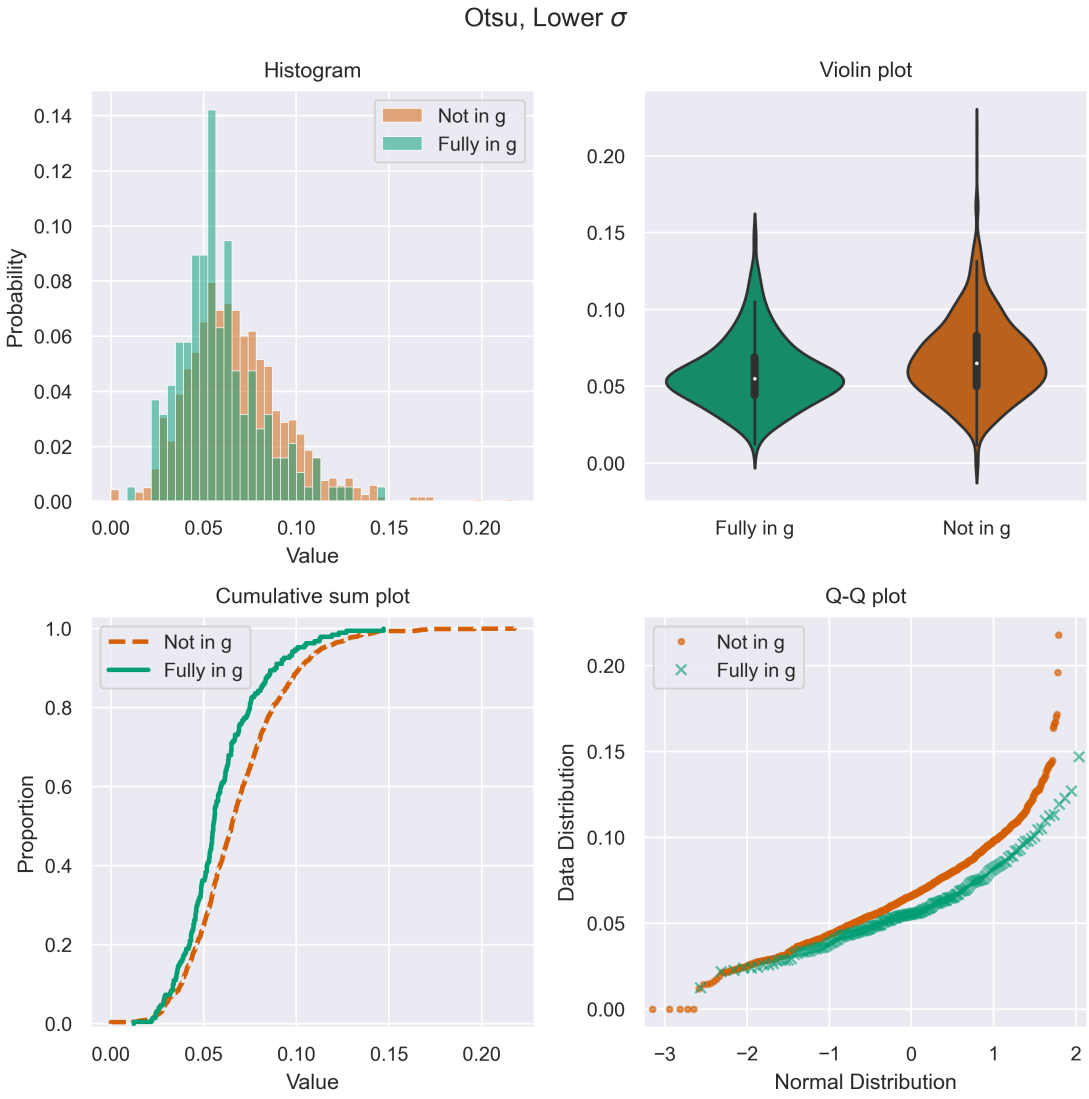


Figure 2.18: Histogram, violin plot, cumulative sum plot and Q-Q plot for the standard deviation of the lower subset defined by the Otsu thresholding algorithm for BALs in the fully-in-g sample and the rest of the BALs.

teristic light-curve of a given sample. This would be an alternative to a pseudo-"composite" light-curve and would allow for a more direct comparison by either visual inspection or clustering in a lower-dimensional representation. This or other ML models have potential in elucidating differences in BAL and non-BAL variability, if any.

Table 2.2: Results of the Kolmogorov-Smirnov and Levene tests when comparing the time-domain features from BALs and non-BALs. The flags indicate the tests indicate there is a significant difference between the given feature between the samples.

Feature	D_{KS}	$p_{val,KS}$	W	$p_{val,W}$	KS Flag	Lev. Flag
Amplitude	0.187	0.000	5.947	0.015		
AndersonDarling	0.119	0.018	0.874	0.350		
Beyond1Std	0.080	0.225	3.172	0.075		
Cusum	0.106	0.047	1.698	0.193		
ExcessVariance	0.103	0.059	2.418	0.120		
Etae	0.145	0.002	0.846	0.358		
InterPercentileRange	0.159	0.000	1.216	0.270		
Kurtosiss	0.054	0.709	0.038	0.846		
LinearFit_slope	0.078	0.261	0.305	0.581		
LinearTrend_slope	0.072	0.353	0.292	0.589		
MagnitudePercRatio	0.096	0.093	4.096	0.043		
MaximumSlope	0.126	0.010	7.347	0.007		✓
MeanVariance	0.174	0.000	2.123	0.145		
MedianAbsDev	0.165	0.000	0.618	0.432		
MedianBuffRangePerc	0.060	0.568	1.237	0.266		
OtsuSplit_diff	0.164	0.000	2.942	0.087		
OtsuSplit_lower	0.213	0.000	8.881	0.003	✓	✓
OtsuSplit_upper	0.166	0.000	3.080	0.080		
PercentAmplitude	0.193	0.000	6.318	0.012		
PercDiffMagPerc	0.168	0.000	1.431	0.232		
Periodogram_peaks	0.119	0.017	0.548	0.459		
Reduced_Chi2	0.135	0.005	8.312	0.004		✓
Skewness	0.111	0.033	1.031	0.310		
StandardDeviation	0.174	0.000	3.416	0.065		
StetsonK	0.075	0.300	0.197	0.658		

Chapter 3

Multimodal Learning Experiments

The task of identifying BAL QSOs through variability is challenging. So far, AGN light-curve classifiers rely on the samples to be processed to have distinct variability behaviors. For instance, even though QSO variability is thought to generally resemble a Damped Random Walk, this is not always the case, or for every AGN type [Kasliwal et al., 2015]. However, in the problem presented here, there is no easily recognizable difference between the light-curves of BAL QSOs and other QSOs, as has been found in other works [Sánchez-Sáez et al., 2018] and was confirmed for our particular dataset in Section 2.4, making this a more complex classification problem.

In this work, we use a MML approach to test its potential for finding BAL QSOs in time-domain surveys such as LSST by building and testing spectrum-assisted light-curve classifiers. In particular, we aim to see if it is possible to connect the shape of the blue-shifted absorption of CIV to variability. MML could potentially allow us to uncover and understand any correlations between the shape of the CIV troughs and variability at a deeper level. We test three different ways of combining these modalities. We test early, late and multiplicative fusion, with an attentive approach whenever possible. We expect the classification done with spectral data or its PCA representation to be more accurate than the one done with light-curves or their features. Even though we do not expect the combined classification to be better than the one done with spectra, here we aim to obtain a light-curve classification that is more accurate than the uni-modal one.

3.1. Data Modalities

We base our work on the sample described in Section 2.1. In this Section, we describe the modality-specific pre-processing of the data done to prepare it for the ML methods.

3.1.1. Spectra

For ML, all the spectra processed should be at the same rest-frame wavelength range. In order to trace the shape of the CIV absorption troughs instead of the whole spectrum of the sources, we select the restframe region $1425\text{Å} \leq \lambda \leq 1600\text{Å}$ in each spectrum, which is equivalent to a redshift selection of $1.671 \leq z \leq 4.759$. Figure 3.1 shows an example spectrum with the selected wavelength range shaded in yellow.

Bad pixels are identified and discarded with the `and_mask` provided by SDSS. Then, the spectra are re-binned to a grid of 503 pixels with bins of 0.346Å such that they all have the same length. This way, we ensure our chopped spectra are all on the same wavelength

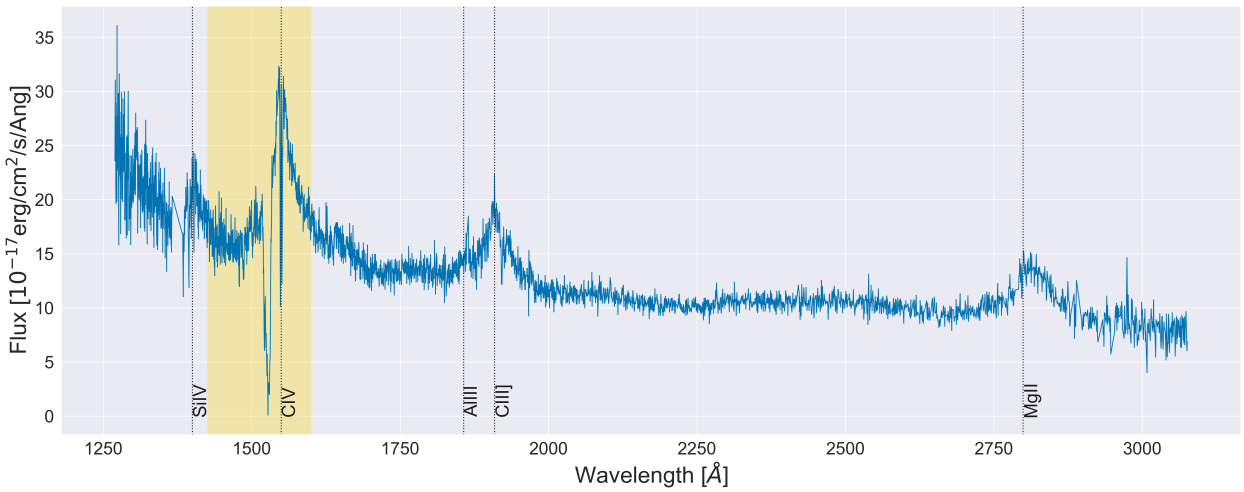


Figure 3.1: Example of spectrum of Hi-BAL SDSS 024304.68+000005.4 at $z = 1.9945$ with the wavelength range used for ML shaded in yellow: $1400 - 1600$Å.

grid. Additionally, fluxes were normalized with the mean flux of each spectrum such that the average is set to 1.0 (see Kao et al. [2024]).

3.1.2. Light-Curves

The light-curves are cleaned and their time-domain features are extracted as described in Section 2.4. The features that describe some aspect of variability (those marked with a $\star$ in the list of time-domain features in Section 2.4.1.2) are used as tabular data, or as a proxy of a lower-dimensional representation of the light-curves. These are 30 features in total.

3.2. Training and Test Sample

In this Section, we describe how the ML sample was selected, built, and separated into the training an test sets.

After the redshift cut done for the treatment of spectra (see Section 3.1.1), our sample was reduced to 843 BALs and 5569 non-BALs. Secondly, we made sure that all of these objects have a present SDSS spectrum. There were two non-BALs and one BAL without a spectrum, which reduced the sample to 842 BALs and 5567 non-BALs. Thirdly, in order to apply multimodal ML, we also require the objects in our ML sample to have a ZTF g-band light-curve available. This requirement further reduced our sample to 5363 non-BALs and 809 BALs. The final ML dataset based on these requirements has two main challenges to be addressed.

Firstly, the number of objects, specially BAL QSOs, is very limited. An important principle in ML is that any model requires a sufficient amount of data in order to learn from it. Over-fitting or under-fitting may occur. The former could lead to poor generalization and low performance, and the latter could result in the model not being able to learn because there is simply not enough to learn from or the model is too simplistic for the given task. An early relevant paper, Banko & Brill [2001] (see also Halevy et al. [2009]) showed that models with varying complexity and size perform just as well when given a sufficiently large amount

of data. However, simply retrieving more data is not always possible. Data augmentation is often used to deal with this problem [e.g. van Dyk & Meng, 2001, Xie et al., 2020]. This consists of creating new modified instances of existing data samples in order to increase its size or diversity. Either way, if possible, retrieving more real data should always be prioritized over more complex techniques.

Secondly, it is quite imbalanced. BAL QSOs encompass only 13% of the sample. Imbalanced datasets are a common problem in ML across many domains and there are extensive methods specialized to deal with this issue [Brownlee, 2021, Gautam & Dey, 2022, Kumar et al., 2021]. Imbalanced datasets can significantly hinder the performance of classification tasks, as algorithms will simply train more times over the majority class, and ignore the minority class. Two common strategies are to over-sample or under-sample the datasets. Over-sampling, much like data augmentation, refers to increasing the number of instances of the minority class to minimize the disparity in size, by either retrieving more data or creating synthetic examples. A popular method used for this is the Synthetic Minority Over-sampling TEchnique (SMOTE) algorithm [Chawla et al., 2002]. The disadvantage of synthetic over-sampling techniques is that it directly depends on the algorithm used, which, if it has some bias or limitation, could affect the performance of the ML model. Under-sampling refers to discard instances from the majority class to make the imbalance less extreme. This can be done either randomly or by some criterion specific to the problem domain (for an example in AGN variability, see Sánchez-Sáez et al. [2021a]). When under-sampling, there is the risk to end up selecting a sub-sample of the majority class that has some bias, which could then affect the final results. It has been found that a combination of over and under-sampling instead of either one on its own tends to be the best approach because it avoids the disadvantages and limitations of each of these techniques. Additionally, the performance of algorithms trained on imbalanced datasets must be adequately evaluated. Only using the accuracy can lead to wrong results: it can easily result in 90% or more because it is correctly classifying the majority class, whilst not providing any good classifications for the minority class. Thus, it is relevant to use appropriate metrics when working with imbalanced datasets. Metrics such as the precision, recall and confusion matrices, should be preferred over the accuracy.

Regarding our imbalanced dataset of QSOs, under-sampling the non-BAL sample is not a good approach because, given the limited number of BALs, it is not ideal to also decrease the overall size of the sample. Data augmentation or synthetic over-sampling are also poor approaches given the multimodal nature of our present task. If we were to apply these methods, we would need to create a synthetic spectrum (or lower dimensional representation of the spectrum) as well as a light-curve (or set of its time-domain features). The algorithm could easily introduce biased correlations between the spectral and time-domain data, which should be avoided since analyzing the intrinsic correlations between these modalities is precisely one of the goals of our work. Therefore, we have opted to over-sample the BAL QSO sample by retrieving more real data. We retrieve 4578 BAL QSOs in the selected redshift range ($1.671 \leq z \leq 4.759$) from the DR16Q catalog built by Lyke et al. [2020] that are not already present in our sample, and download and process their spectra and light-curves as described in Sections 2.3 and 2.4. Their spectra were recovered for all BAL QSOs, and their light-curves for 4342 of them. Overall, we end up with a BAL QSO sample of 5151 instances, in the required redshift range and with both data modalities available. A possible bias that the extra BAL QSO sample could introduce is that these were not selected according to the same criteria as the original dataset.

Our final ML dataset has 5151 BAL QSOs and 5363 non-BAL QSOs. It is no longer

imbalanced, as BALs encompass 48.99% of the sample. To continue to take advantage of the cleanliness of our original sample, we use its 809 BALs, (15.7% of the ML sample) for testing and validating the performance of the models, whilst we use the retrieved BALs from the DR16Q for training. We randomly split the non-BALs at the same fraction for training and testing.

3.3. Multimodal Learning Methods

In this Section, we present the ML experiments done. We describe two multimodal ensembles: one with tabular models and other with dense neural networks (NNs) on tabular data.

To evaluate the performance of our models, we look into the accuracy, precision, recall and F_1 scores. They are defined with the number of True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN):

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.3}$$

$$F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}} \tag{3.4}$$

While the accuracy gives an overall measure of correct predictions, the precision and recall are particularly useful in our present work. The former quantifies the ability of the model to avoid labeling negative instances as positive, which would mean labeling non-BALs as BALs. Thus, the precision measures the purity of the predicted BAL sample. On the other hand, the recall gives a measure of the models ability to find all positive instances, or labeling all BALs correctly, providing a measure of completeness. A particularly useful metric is the F_1 score. It consists of the harmonic mean between the precision and the recall, and concisely provides an overall measure of purity and completeness in a single metric. It ranges from zero to one, where higher values indicate better performance.

Additionally, the confusion matrix is a useful visual tool, consisting of a table that compares true and predicted labels. The Receiver Operating Characteristic (ROC) curve and the area under it (Area Under the Curve, AUC) are also a good visual measure for the performance of the model. They plot FP rates against TP rates: the closest the curve passes to the upper left corner of the plot, the better the performance of the model. Similarly, a larger AUC indicates better predictions.

3.3.1. Tree Ensemble Models

Working with tabular data and traditional ML is a powerful approach [Shwartz-Ziv & Armon, 2021]. Random Forests (RFs) tend to even outperform quite complex DL models. Here, we use tree ensemble models on the tabular or 1-dimensional representation of each of our data modalities. In particular, we test the performance of RFs and extreme gradient

boosting (XGBoost).

RFs [Breiman, 2001, Kan Ho, 2016] consist of an ensemble of multiple decision trees. Its key advantage is its double source of randomness in training to avoid over-fitting. Each tree is trained on a random sub-sample of the training data selected by bootstrap aggregating (bagging). Additionally, each node of each tree is trained on a random subset of the inputted features, and the best split is chosen among those features only. Then, the trees are fused by averaging them. RFs have many practical advantages. Even in high-dimensional datasets with many irrelevant features, they are capable of finding the most relevant ones, which is retrievable in the feature importance measure they provide. They are also quite robust to outliers and train efficiently. This algorithm also extrapolates well to work on imbalanced datasets [Amrehn et al., 2019]. The balanced RF implemented by `imblearn` [Lemaître et al., 2017] does this by drawing a random sub-sample from the majority class that is the same size as the minority class bootstrap-selected sample, ensuring balanced training in each tree, whilst still being able to see all instances of the data.

Furthermore, XGBoost is also a powerful method [Chen & Guestrin, 2016]. It consists of an ensemble of weak decision trees that results in a strong predictor by learning in a sequential manner, where each tree attempts to correct the errors from the previous one. It applies gradient descent to minimize the given loss function, i.e. to iteratively move in the direction of the steepest descent in order to minimize the given function in as little iterations as possible. XGBoost in particular is called "extreme" because of its several improvements over other gradient boosting methods. For instance, it prevents over-fitting by implementing regularization terms, is computationally efficient, systematically handles missing data and supports early stopping.

For both RFs and XGBoost, we ran a grid search in order to tune their hyper-parameters with a K-fold cross-validation $k = 5$ folds, and using the recall as the target score to prioritize the completeness over the purity. For the RFs, we set use the out-of-bag (oob) samples to true (i.e. using the unseen data instances at each tree to validate training). For both models [Probst et al., 2019], we fit the number of estimators (i.e. the number of decision trees in the RF) and the maximum depth (i.e. the maximum number of nodes in each tree). The rest of the hyper-parameters were left to the default values implemented in `scikit-learn`, including the Gini split criterion.

3.3.1.1. Spectral Dimensionality Reduction and Tabular Models

A useful technique for the classification of BAL QSO spectra is to reduce their dimensionality before feeding them to an algorithm. Here, we use a similar approach to Kao et al. [2024] in order to test which combination of dimensionality reduction technique and tree ensemble classifier works best for our present dataset. The methods we use for dimensionality reduction are described bellow.

Principal Component Analysis

PCA Jolliffe & Cadima [2016] is likely the most popular dimensionality reduction method. In simply terms, it works by looking for the hyperplane that can preserve the highest variance, and then projecting the data onto it. It has the advantage that it is possible to find the optimal number of PCA components in a methodical way based on the variance. A good criterion is to make sure that $\geq 90\%$ of it is preserved.

We found that for our dataset, 236 PCA components preserve 90.05% of the variance. The best RF has a maximum depth of 50 nodes and 100 trees. The best XGBoost has a maximum depth of 3 nodes and 500 trees.

Locally Linear Embedding

Locally Linear Embedding (LLE) Ghojogh et al. [2020] is a manifold method that learns local symmetries by linearly modeling local relations between the x-nearest neighbors, and looks for a lower-dimensional representation that best preserves them.

We chose the best possible number of dimensions and neighbors by doing a grid search on these parameters, with a K-fold cross-validation with $k = 5$, and with a target score of the mean square error. We found that the best LLE representation for our spectra has five components with 5-nearest neighbors. The best RF on this data has a maximum depth of 50 nodes and 500 trees. The best XGBoost has a maximum depth of 3 nodes and 100 trees.

Uniform Manifold Approximation and Projection

This manifold method was developed by McInnes et al. [2020] and it is based on Riemann geometry and algebraic topology [9]. It is faster and more robust toward large datasets than other methods such as t-SNE [van der Maaten & Hinton, 2008]. The relevant hyperparameters in UMAP are the number of approximate nearest neighbors, which determines whether the algorithm focuses more on the local or global structure of the data, and the minimum distance between points in the low-dimensional space.

Once again, we found the best possible number of dimensions, nearest neighbors and minimum distance with a 5-fold cross-validation grid search, and with a target score of the mean square error. The best UMAP representation has three components, 5-nearest neighbors and a minimum distance of 0.1. The best RF on this data has a maximum depth of 50 nodes and 1000 trees. The best XGBoost has a maximum depth of 3 nodes and 100 trees.

Result

The accuracy, precision and recall scores obtained for each of the best models described here are summarized in Table 3.1. The best results obtained are obtained when using PCA. This is also the computationally cheapest method. By preserving 90% of the variance, we lowered the dimensionality of our data from 503 to 242 components. When applying LLE and UMAP, we obtained similar accuracies and, in spite of having higher precision scores and a significantly lower number of components, the recall is 5% to 10% lower. Furthermore, the choice between using a RF or XGBoost on the PCA representation is less obvious, given that the recall is higher for the former but precision is higher for the latter. We choose the RF over XGBoost to prioritize the completeness of the BAL predicted sample over its purity, but note that this is a subjective assessment, since overall the performance of both models is fairly good.

Moreover, we note that by applying a similar approach, Kao et al. [2024] also found a combination of PCA and RF or XGBoost to be the best performing technique for the dimensionality reduction and classification of BAL QSO spectra. PCA has also been found to be a good dimensionality reduction technique for SDSS galaxy and QSO spectra in general, with applications implemented in the `astroML` package[10] [Ivezić et al., 2014, Vanderplas et al., 2012], and Brodzeller & Dawson [2022] also use PCA to model QSO spectra from SDSS.

[9] See https://pair-code.github.io/understanding-umap/
[10] See https://www.astroml.org/book_figures_1ed/chapter7/fig_PCA_LLE.html

Table 3.1: Results of spectral dimensionality reduction done for spectra per method, number of dimensions, classifier and its accuracy, precision, recall and F_1 scores.

Method	Dims.	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F_1
PCA	236	RF	87.00	88.06	**85.66**	0.867
		XGB	**87.86**	**92.16**	82.82	**0.872**
LLE	5	RF	83.16	89.43	75.28	0.817
		XGB	83.59	90.60	75.03	0.821
UMAP	3	RF	85.82	91.43	79.11	0.848
		XGB	85.08	91.64	77.26	0.838

3.3.1.2. Light-Curve Tabular Models

We apply the grid searches directly on the 30 light-curve features selected for ML (see Section 3.1.2; note that the errors, noise and χ^2 of the `LinearFit` and `LinearTrend` were excluded). The best RF has a maximum depth of 20 nodes and 100 trees, and the best XGBoost model has a maximum depth of 3 nodes and 500 trees.

Table 3.2 shows the results for the best classifiers. The RF performs better than the XGBoost model. However, its performance is quite poor. This behavior was expected, as it was seen in Section 2.4 that there is little difference between the features of BALs and non-BALs. The recall score is particularly low, which indicates that the correctly classified BAL QSOs are a minority.

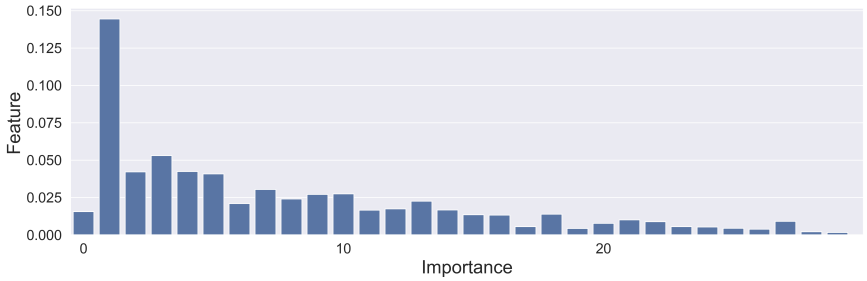
Table 3.2: Results of the light-curve features classification by model and its accuracy, precision, recall and F_1 scores

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F_1-score
RF	**55.01**	**69.94**	**14.07**	**0.234**
XGB	53.74	66.42	10.99	0.189

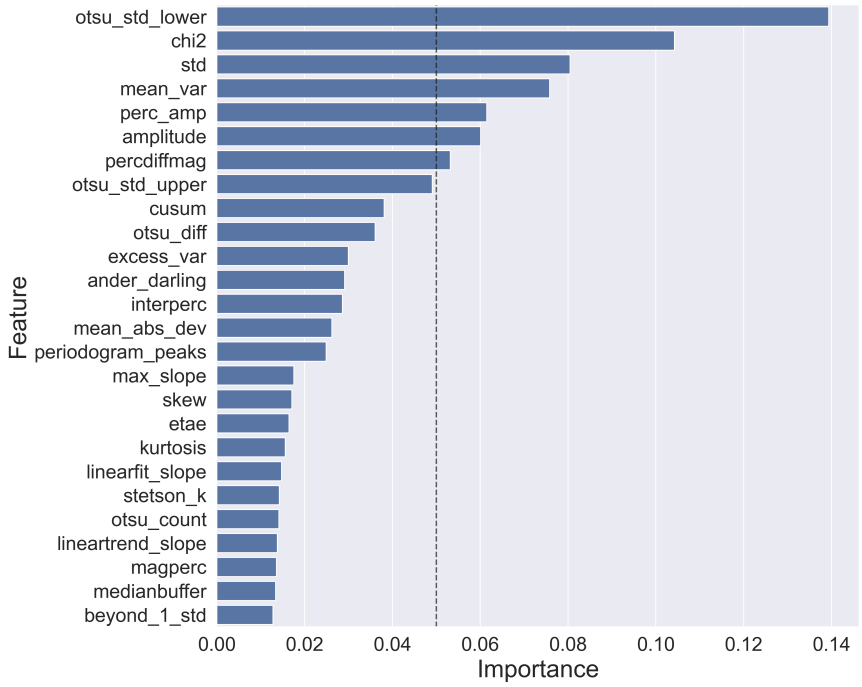
3.3.1.3. Multimodal Random Forests

Before combining the decision probabilities, we inspect the feature importances of each RF. This tells us which features were the most relevant in reaching the final decisions. The most important features are those that, when used to split a node, result in the largest Gini impurity decrease. Figure 3.2.a shows the importances for the spectral RF, and Figure 3.2.b the ones for the light-curve RF. For the spectral RF, we see that there is only a handful of PCA components with importance higher than 10% and even 5%. Therefore, for the multimodal RF we will only use the first 15 PCA components. Usually, using the largest number of features possible is the best option. However, when too many features have such little importance, they can confuse the model. For the light-curve RF, the behavior is similar. Within the time-domain features, we choose those features with importance higher than 5%.

Figure 3.3 displays the ROC curves and confusion matrices for the separate RFs computed on the sub-sample of features selected by their importance. As expected, the classification obtained with the spectral PCA components is significantly better.



(a) Feature importances of the RF trained on the PCA representation of spectra. Only the first 30 PCA components are shown, out of 236.



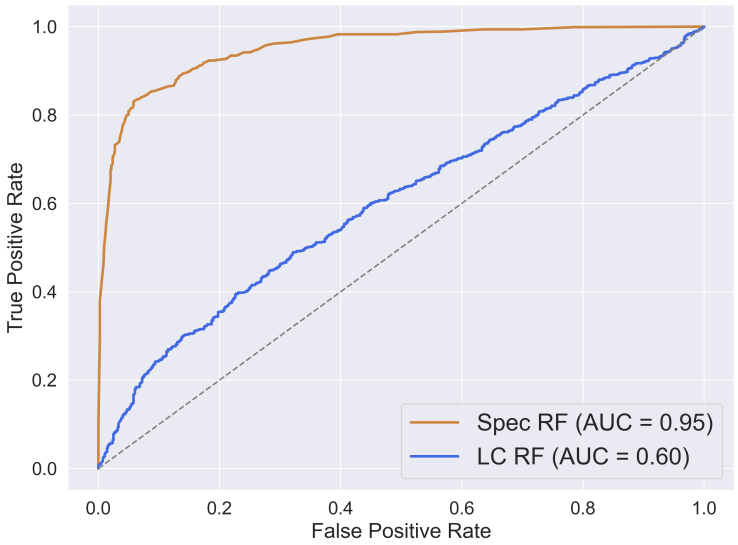
(b) Feature importances of the RF trained on the light-curve features.

Figure 3.2: Feature importances of the uni-modal RFs.

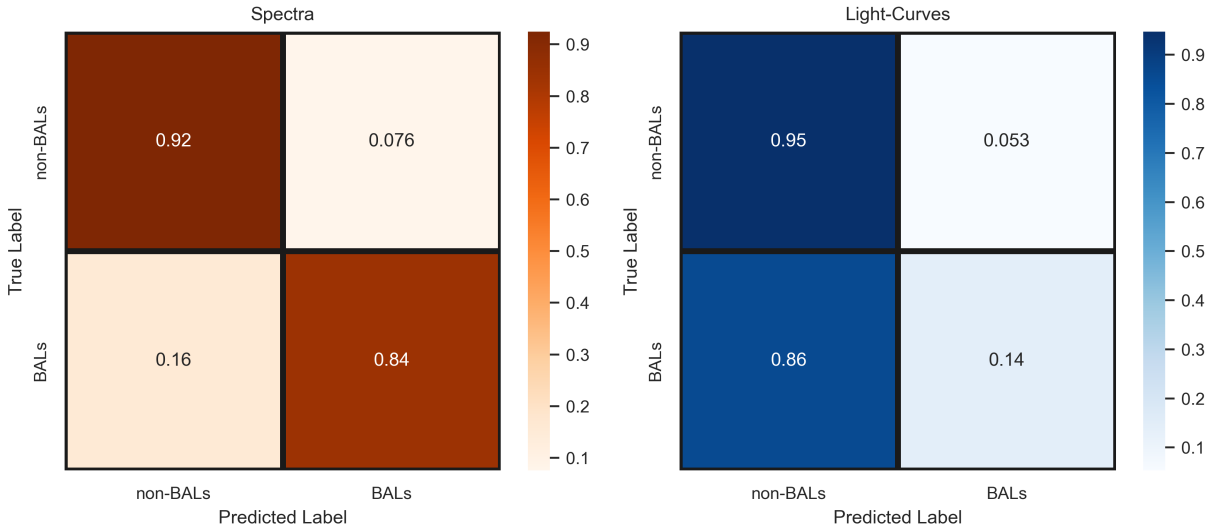
Late Fusion

Next, we test the performance of an ensemble of the best found modality-specific models mentioned above. For this, we compute a weighted average of the classification probabilities. The weights are calculated as follows: first, the prediction with higher maximum probability, which indicates higher confidence in the prediction, is chosen for each QSO; then, they are normalized by the sum of the confidence levels from both classifiers, ensuring the weights sum up to one. Given that the spectral classification is more reliable, this will naturally favor it over the light-curve classification predictions.

The resulting prediction has an accuracy of 81.56%, a precision of 94.84%, a recall of 65.88% and an F_1 score of 0.778. Figure 3.4 displays the obtained confusion matrix. When comparing with the matrices in Figure 3.3.b, we see that the ensemble is much more reliable at finding BAL QSOs than the light-curve classifier alone. Indeed, the recall score improves by 51.79%. As a first test, this result indicates there is significant potential for MML in the present task.



(a) ROC curves of the spectral and light-curve RFs.



(b) Confusion matrices for the spectral and light-curve classifications.

Figure 3.3: Results of the uni-modal RFs.

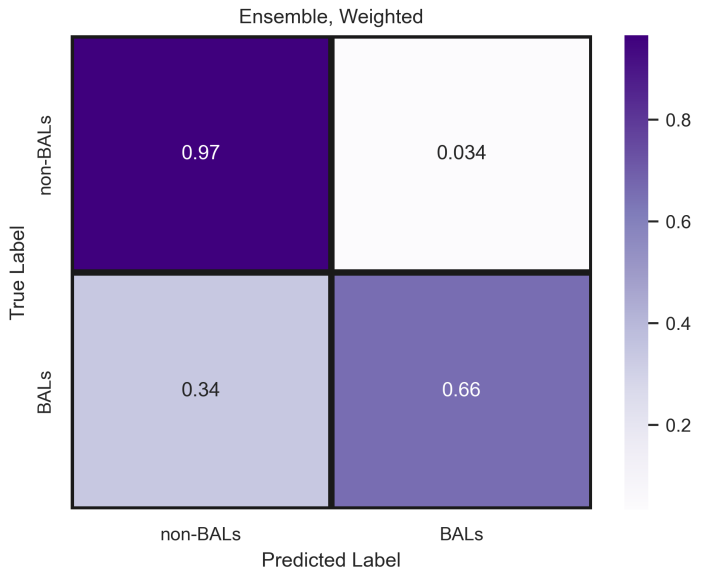


Figure 3.4: Confusion matrix of the late-fused RFs.

Early Fusion

As discussed in the introduction (see Section 1.3.1), late fusion allows for separate treatment of the modalities. However, early fusion is the appropriate technique when we wish to look into the deeper-level correlations between the modalities. Here, we train a single RF on a joint representation of the spectral PCA components and light-curve features. The important decision to be made here is how exactly to merge them such that the classifier can properly process them. Here, we concatenate them and scale them with the `RobustScaler` implemented by `scikit-learn`, which, as its name implies, is robust against outliers: it subtracts the median $\tilde{X}$ from each value and divides by the interquartile range (IQR), which is the range between the 25th and 75th percentiles [11]:

$$X_{scaled} = \frac{\left(X - \tilde{X}\right)}{\mathrm{IQR}(X)} \tag{3.5}$$

We looked for the best classifier on the concatenated data by running a grid search on a RF and an XGBoost model, in the same way as described in 3.3.1. The best RF and XGBoost models have 500 and 300 trees, with a maximum of 20 and 3 nodes respectively. Their performance is shown in Table 3.3. We choose the XGBoost model given that its performance is slightly better.

Table 3.3: Results of the classification on the early-fused concatenated features by model and its accuracy, precision, recall and F_1 scores

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F_1-score
RF	83.68	93.54	71.57	0.811
XGB	**84.40**	**94.94**	**71.94**	**0.819**

[11] See also https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html

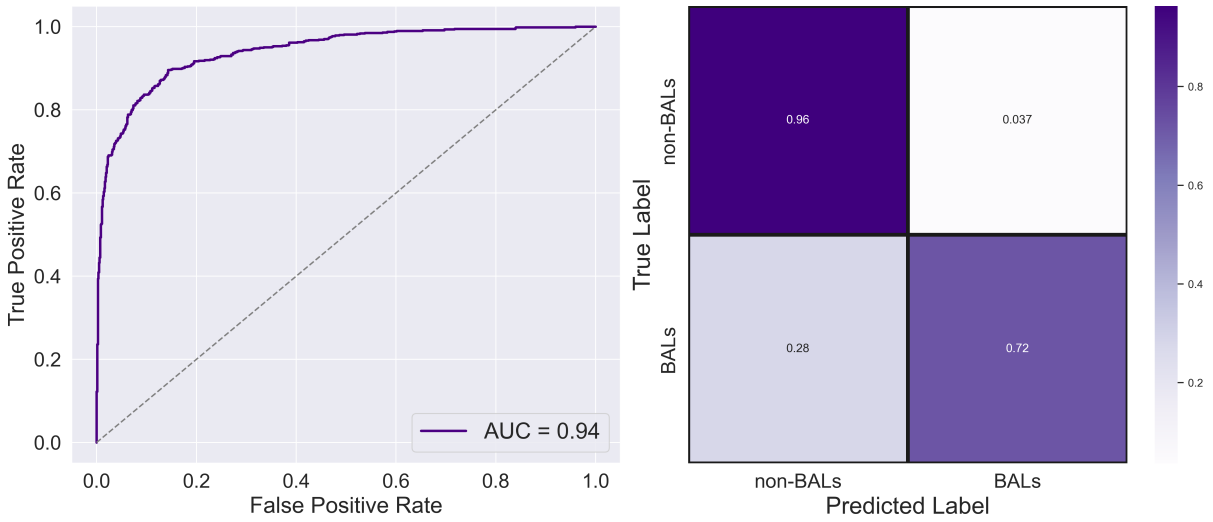


Figure 3.5: ROC curve and confusion matrix of the early-fused RF.

The results of the early-fused RF are displayed in Figure 3.5. We see that the AUC is equal to 0.93, which is quite good. All the evaluated metrics are better that those obtained by the RF built with late fusion, even if by a minimal difference. In particular, the amount of BALs that are correctly classified increase from 66% to 72/%, as seen in the confusion matrices and the increase of 6.06% in the recall score. Therefore, when using tabular data, we conclude that early fusion is preferable in this case.

Furthermore, it is relevant to see how the model is learning from both modalities, and if it finds any correlations between them. This can be done by inspecting the feature importance of the model. In Figure 3.6, we see that the few most important features are indeed a mixture of both modalities and not from a single one. Interestingly, the most relevant input with an importance of 32.82% is a light-curve feature, the standard deviation of the lower subset defined by the Otsu thresholding algorithm [Otsu, 1979]. This can be interpreted as the standard deviation of the baseline variability. After it, the second PCA component is the most important feature with an importance of 20.22%. This indicates that the model indeed benefits from the combined modalities, showing promise for this approach.

3.3.2. Dense Neural Network

The next multimodal test we run is a dense NN trained on the same tabular data as the early-fused model described in the previous section.

3.3.2.1. Description of the Model and Fusion Technique

Figure 3.7 displays the structure of the model. First, each modality is processed separately through two dense layers with ReLU activation functions, with a batch normalization and dropout layers and L_2 regularization at each of them to prevent over-fitting. The former maintains the mean and standard deviation close to 0 and 1 [Ioffe & Szegedy, 2015], and the latter randomly turns units off such that there is not a significant imbalance between over-trained and under-trained neurons [Labach et al., 2019, Srivastava et al., 2014]; we use a dropout fraction of 20%. The spectral PCA components are increased in shape in the NN from 15 to 32 and 64, whilst the light-curve features go from seven to 16 and 32. Then,

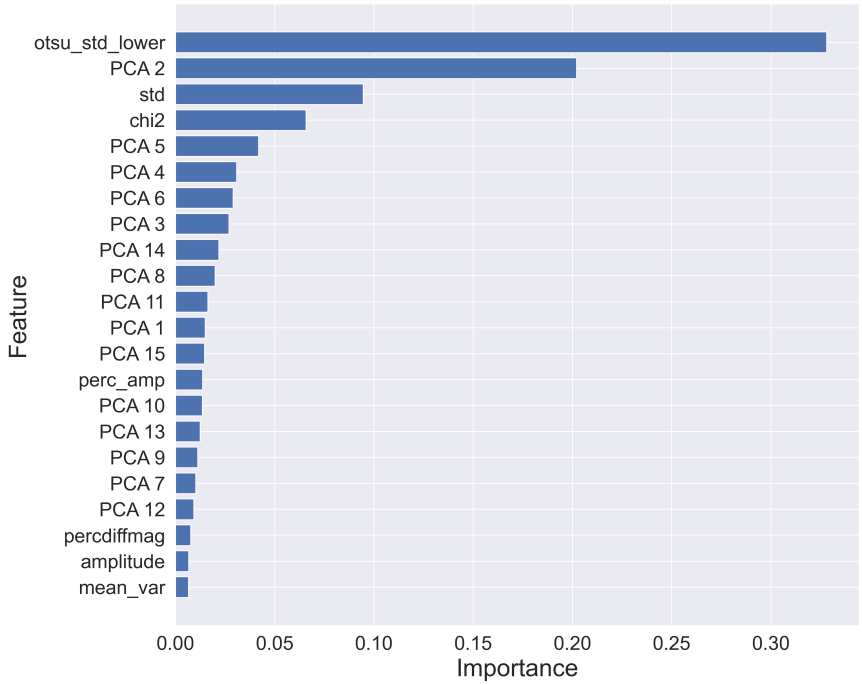


Figure 3.6: Feature importance of the early-fused RF.

there is a third dense layer with a sigmoid activation that learns to predict how reliable the output of each modality is. Its output is used by an addition and subsequent lambda layers simply to normalize the sum of the reliability scores to unity. The weights have an added minimum of 10^{-6} to ensure there is no division by zero. Next, we apply multiplicative fusion by weighting each modality by their corresponding reliability scores, and then concatenating them. Then, the concatenated data are passed through an additional dense layer (with batch normalization, dropout and ReLU activation as well) and then are finally outputted by the final dense layer with a soft-max activation.

When training the model, we use an Adam optimizer [Kingma & Ba, 2017] with a learning rate of 10^{-3}, categorical cross-entropy for the loss function, and accuracy for the target metric. We also apply an early stopping regularization with a patience of 15 epochs, which means the training will stop after 15 epochs with no improvement, and we reduce the learning rate by a factor of 0.5 when one of the target metrics has become stagnant for 5 epochs. Additionally, we use the same fraction of train to test sample sizes to define a validation set from the training data.

3.3.2.2. Results

Figure 3.8 displays the results from the dense NN. The model trained for 54 epochs before stopping. The top panel shows the loss and accuracy for the training and validation sets, which behave as expected with no major signs of over-fitting. Before epoch ~ 30, the validation set exhibits a zig-zag pattern which could potentially indicate issues in training such as over-fitting. However, this behavior subsides later on, most likely indicating that the several regularization techniques implemented were effective in preventing over-fitting. In the bottom panel of the figure, we display the ROC curve, which is quite good, with an AUC of 0.96, and the confusion matrix. We see that, even though the improvement is not drastic, the NN predictions are more reliable over the ones by the early-fused XGBoost model (see

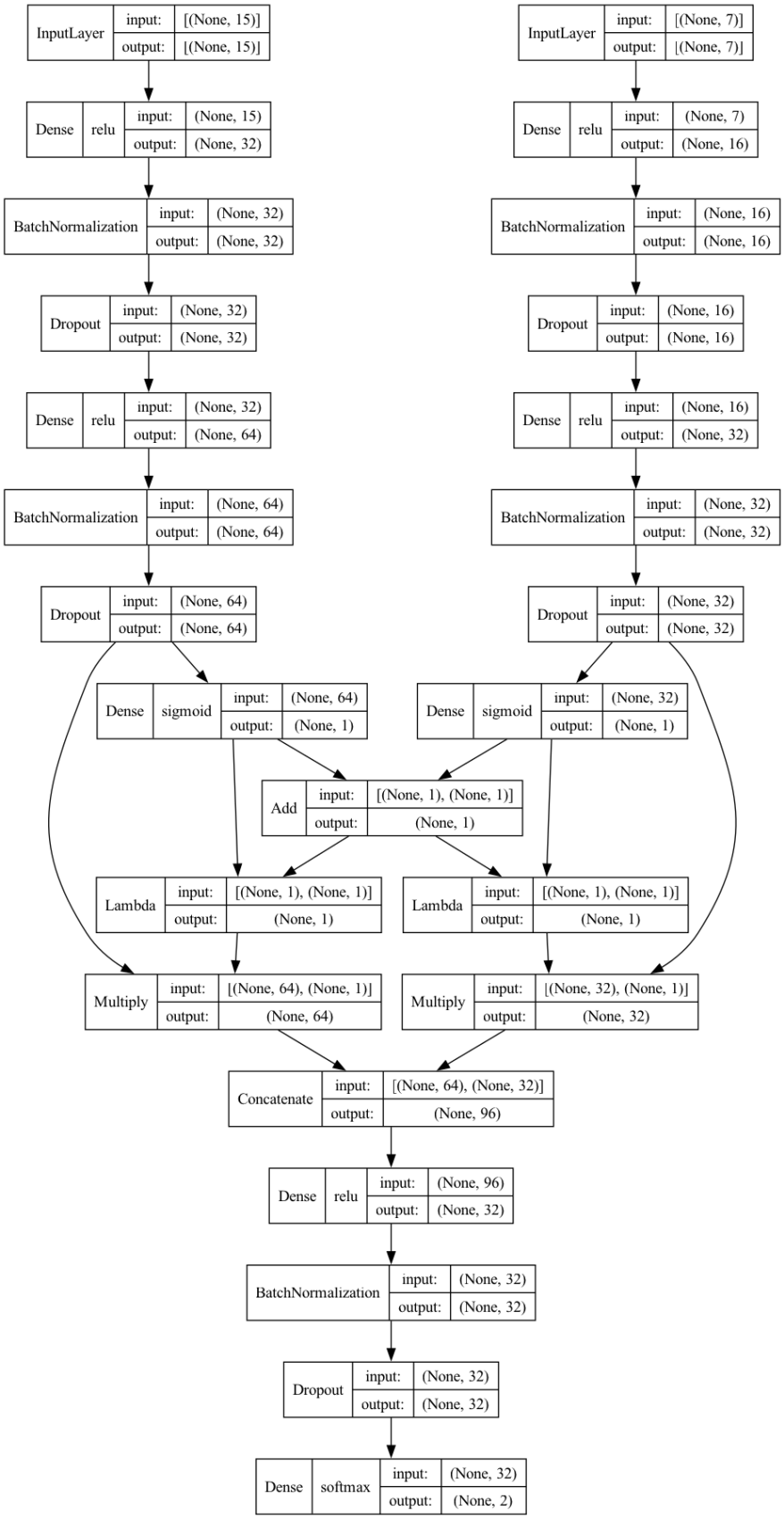
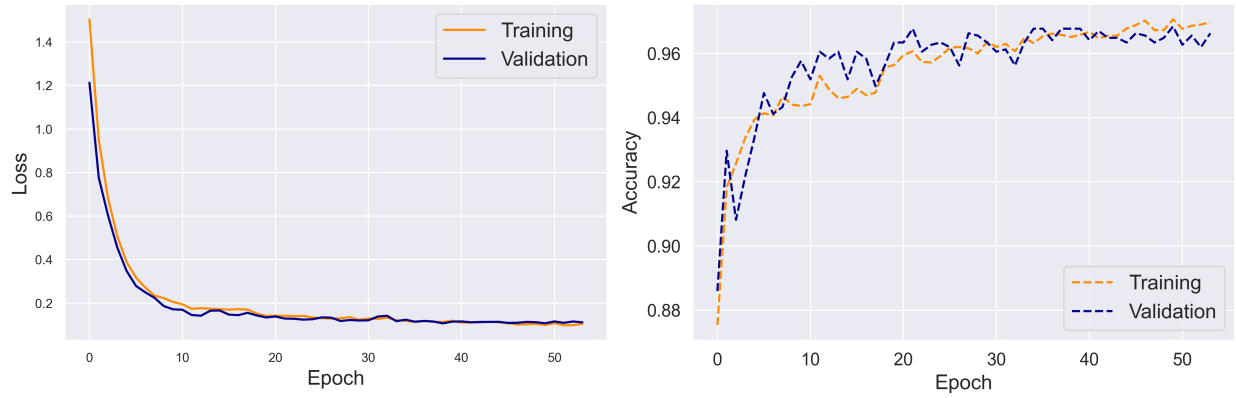
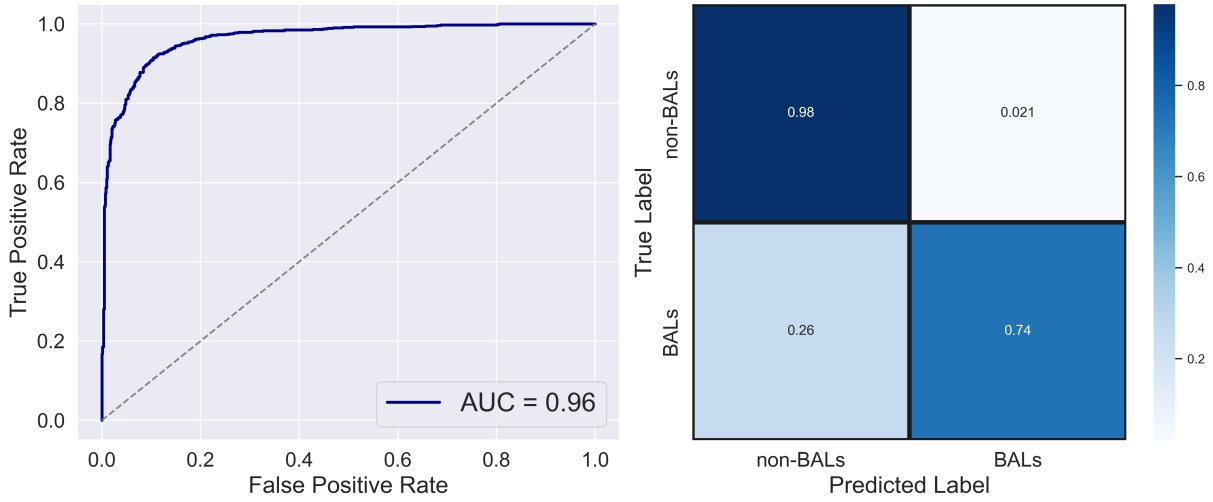


Figure 3.7: Structure of the dense NN trained on both modalities with multiplicative and attentive fusion.



(a) Loss and accuracy of the model by epoch.



(b) Confusion matrices for the spectral and light-curve classifications.

Figure 3.8: Results of the uni-modal RFs.

Figure 3.5). The model has an overall accuracy of 86.03%, and weighted precision, recall and F_1-score of 88.10%, 86.03% and 0.858 respectively ("weighted" here means that the averages were weighted by the number of true positives in each class). The non-weighted precision, recall and F_1-scores are equal to 97.07%, 73.67% and 0.838 respectively. Compared with the early-fused XGBoost model, we see an improvement of 2-3% in each of the evaluated metrics. In particular, the weighted recall is fairly good, indicating that, so far, the dense NN is the best at recovering the largest possible amount of correctly classified BAL QSOs.

Chapter 4

Summary and Future Prospects

In this work, we presented a comprehensive analysis of a clean sample consisting of 1419 BAL QSOs and 41086 non-BAL QSOs built by Naddaf et al. [2023] from the SDSS DR7 QSO catalogue [Shen et al., 2011].

Our first aim was to provide a detailed characterization of this sample. We did this by constructing mean SEDs, composite spectra and analyzing their light-curves. We compare the BAL QSOs to the non-BAL QSOs through these tools. Additionally, we define the "fully-in-g" sub-sample, which corresponds to those BAL QSOs which CIV absorption troughs land fully within the g-band of SDSS. We compare the objects in this sub-sample to the rest of the BAL QSOs (called the "not-in-g" sample) to see if the position of the absorption features and, in particular, the variability of the CIV troughs can have an impact on the overall photometry, SED or light-curves of the sample.

Our derived mean SEDs successfully recover the redder UV-to-optical continuum found in previous works, as well as a steeper slope in the IR range. Both SED characteristics point to high dust extinction in the BAL QSOs. Our results are consistent with dust components being present in the outflows, which in turn is thought to be a major factor in the accelerations mechanisms of the moving material, particularly radiation pressure. We also find the SED derived for high-luminosity BAL QSOs by Saccheo et al. [2023] has a flatter UV continuum, but discard that this difference is a function of luminosity given that several of our BAL QSOs have high luminosities as well. In addition, the derived SED for the non-BALs in our sample is similar to others found in the literature, in particular the one by Krawczyk et al. [2013]. We explain the seen differences by a different redshift distribution, sample sizes or other selection effects. We also find that the SED for the fully-in-g sample has a more pronounced dip, which could potentially indicate stronger dust extinction. However, given the limited number of objects used to derive this mean SED, it is not possible to conclude this physical difference. Further studies with larger samples in the given redshift ranges should be conducted to see if this behavior is factual or if it is just a selection effect of the sample studied here.

Furthermore, we look at the spectroscopic characteristics of the sample, based on spectra fetched from SDSS DR18. We also fetch the Balnicity and absorption indices (BI and AI) from the DR16Q [Lyke et al., 2020]. CIV emission is similar between BAL and non-BAL QSOs, indicating that what distincts them is indeed the blueshifted troughs. Within BAL QSOs, more extreme absorption is less common.

We build the composite spectra in order intuitively describe the characteristics of the overall sample. We are able to recover the distinct spectral shape of Hi-BALs, Lo-BALs and FeLo-BALs. We see the redder continuum in the Lo-BAL composite, but fail to see significant

absorption blueward of the MgII line, which is most likely due to varying trough shapes that wash each other out. Moreover, in the FeLo-BAL composite, the rarest class, we are able to see their defining absorptions at FeII lines, as well as excess emission at several regions of its continuum, which is not seen in the other BAL classes. FeLo-BALs are a distinct and special case with dedicated studies in the literature, and in our composites we are able to see their unique characteristics. Moreover, we built separate composites separating our BAL QSO sample by BI bins. Consistent with stronger dust extinction, we find that for higher BI, the continuum of the composite tends to be flatter. We are also able to see varying absorption shapes, with different depths, widths and blue-shifts, without any trend with BI. Indeed, as the BI indicates any absorption and does not differentiate by properties, we propose that creating a more detailed description of the CIV absorption by its shape, depth and blue-shift will be key in characterizing the structure, distribution and dynamics of the outflowing material in BAL QSOs and their relationship to the host galaxies, which would be essential for future AGN feedback studies.

Then, we looked into the variability of our sample by studying their ZTF g-band light-curves. We computed their time-domain features with code used by several ZTF and LSST brokers. BAL QSOs are thought not to have any characteristic variability behaciour that can easily distinguish them from other QSOs. To confirm whether this is true for our present sample, we conducted statistical tests to compare the distributions of those light-curve features that can be interpreted as some aspect of the variability of BALs and non-BALs. We also compare the features of BAL QSOs in the fully-in-g and not-in-g sample, which is consistent given that the g-band in ZTF and SDSS cover a similar wavelength range. We found only slight differences in a couple of features, and no significantly distinct variability behavior attributed solely to BAL QSOs as opposed to non-BAL QSOs. Some of the objects in our sample have negative excess variance, indicating that they could possibly be not variable at all. Moreover, when comparing the features of the BALs in the fully-in-g and not-in-g samples, we find a small difference in the distributions of the standard deviation of the lower subset defined by the Otsu thresholding algorithm, which can be interpreted as the standard deviation of the baseline variability. However, this difference is not hugely significant, making it not plausible to conclude that this is the key for BAL QSO identification via variability.

The second main aim of this work was to test whethter a multimodal learning approach can assist in the identification of BAL QSOs through variability. We motivate this with the LSST in mind. Finding a way to find these objects through their light-curves will be key in order to not miss the large amount of them expected to be found, and to use them for invaluable AGN feedback and galaxy evolution studies. The modalities we work with here are the SDSS spectra restricted to $1425\text{Å} \leq \lambda \leq 1600\text{Å}$ in the restframe, which covers the CIV absorptions well, and the clean ZTF-g light-curves. The sample of our data that satisfies these requirements is far too small and imbalanced for ML (5363 non-BALs and 809 BALs). In order to avoid any systematic biases, we choose to gather more real data instead of applying data augmentation algorithms. We recover 4342 additional BAL QSOs with both SDSS spectra and ZTF light-curves, such that our ML sample is no longer imbalanced and both classes have more than 5100 instances. We use the BAL QSOs in the originally defined sample for the test set in order to continue to benefit from the cleanliness of this sample selection. We test two MML models: tree ensembles and a dense NN, both trained on tabular data.

We process the spectra and light-curves separately at first. We look for the best combination between a method to reduce the dimensionality of the spectra and classify their new

representation. We find that PCA is the best at representing spectra at a lower dimension, which has been found and applied by other works as well. Both a random forest (RF) and extreme gradient boosting (XGBoost) perform similarly well, with the RF obtainind a better completeness for the correctly classified BAL QSOs. Similarly, we test whether a RF or a XGBoost model is better at classifying light-curve features. Here, we use only those that are interpretable as some aspect of variability and will not introduce any bias. We find that the RF has a better performance than the XGBoost model, but it is still fairly poor, which is expected given that there was no significant difference found between BAL and non-BAL QSO features or variability characteristics.

Then, we test two multimodal RF: one with early fusion (i.e. concatenation of the features of both modalities before inputting them into the model) and late fusion (i.e. decision level fusion where the prediction probabilities are combined by, for instance, averaging them). For the late-fused RFs, we apply a weighted average of the decision probabilities that allows for the model to automatically pay more attention to the modality that is more reliable, i.e. the spectra. For the early-fused model, we take an extra step to scale the concatenated features so that they can be consistently processed, and then we look for the best performance between a RF and an XGBoost model, and find that the latter provides slightly better predictions. Finally, we compared the performance of the early-fused and late-fused models, and found that early-fusion is preferable given that it is able to learn correlations between the modalities at a deeper level. Indeed, the feature importances of the early-fused XGBoost model reveal that the most relevant features are not from a single modality, but rather a combination from both.

We note that, interestingly, the most relevant feature is the standard deviation of the lower subset defined by the Otsu thresholding algorithm, which was found to have one of the most noticeable discrepancies between BALs in the fully-in-g and not-in-g samples, as seen in Chapter 2. To rule out this being a coincidence or a bias, further studies looking into the Otsu thresholding algorithm and its significance for variability classification of BAL QSOs should be conducted.

The second model we build is a dense NN with multiplicative and attentive fusion. The former ensures that the model will learn deep inter-modality correlations, and the latter that the most reliable modality has larger weights. We implement several techniques to prevent over-fitting and see that the NN trains well across epochs and successfully provides the best predictions out of all the MML models tested here. It was able to correctly identify 74% of the BAL QSOs in the test set, as opposed to only 14% obtained by light-curve feature classification on its own. We conclude that multimodality shows great potential for the task of identifying BAL QSOs through variability.

Given the promising results from our multimodal approach, future studies should focus on refining these models further. A regressive MML model able to predict spectral features from light-curve ones could prove useful. We also propose the potential use of a Variational Auto-Encoder (VAEs) [e.g. Zhao et al., 2017] accompanied by such a model. VAEs have been found to be able to successfully learn latent representations of spectra, and they have the key advantage of providing synthetic spectra drawn from the latent space. Thus, we propose the implementation of a MML model built on light-curve features and latent representations of spectra obtained by a dedicated VAE. Implementing the following pipeline for the LSST could be ground-breaking: LSST light-curves will provide features that can be processed by the MML model, which in turn will be built to predict the corresponding spectral latent representations; then, the latent features could be inputted to the VAE, which can then pro-

duce synthetic spectra from the latent space. This framework has the potential for enabling discoveries that are currently not possible in BAL QSO studies.

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from tensorflow.org

Akkus, C., Chu, L., Djakovic, V., et al. 2022, Multimodal Deep Learning

Alegre, L., Best, P., Sabater, J., et al. 2024, Identification of multi-component LOFAR sources with multi-modal deep learning, arXiv:2405.18584 [astro-ph]

Alexander, E. Date of access: 2024, Unified model of AGN, https://emmaalexander.github.io/images/unified_agn.png

Allevato, V., Paolillo, M., Papadakis, I., & Pinto, C. 2013, The Astrophysical Journal, 771, 9

Almeida, A., Anderson, S. F., Argudo-Fernández, M., et al. 2023, The Astrophysical Journal Supplement Series, 267, 44, aDS Bibcode: 2023ApJS..267...44A

Amrehn, M., Mualla, F., Angelopoulou, E., Steidl, S., & Maier, A. 2019, The Random Forest Classifier in WEKA: Discussion and New Developments for Imbalanced Data

Antonucci, R. 1993, Annual Review of Astronomy and Astrophysics, 31, 473

Antonucci, R. R. J. & Miller, J. S. 1985, The Astrophysical Journal, 297, 621

Arav, N., Borguet, B., Chamberlain, C., Edmonds, D., & Danforth, C. 2013, Monthly Notices of the Royal Astronomical Society, 436, 3286

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. 2017, Multimodal Machine Learning: A Survey and Taxonomy, arXiv:1705.09406 [cs]

Banko, M. & Brill, E. 2001, in Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, ACL '01 (USA: Association for Computational Linguistics), 26–33

Baron, D. & Poznanski, D. 2017, Monthly Notices of the Royal Astronomical Society, 465, 4530, arXiv:1611.07526 [astro-ph]

Berk, D. E. V., Shen, J., Yip, C.-W., et al. 2006, The Astronomical Journal, 131, 84

Bessiere, P. S., Almeida, C. R., Holden, L. R., Tadhunter, C. N., & Canalizo, G. 2024, QSOFEED: The relationship between star formation and AGN Feedback, arXiv:2405.06421 [astro-ph]

Bischetti, M., Feruglio, C., D'Odorico, V., et al. 2022, Nature, 605, 244

Bischetti, M., Fiore, F., Feruglio, C., et al. 2023, The Astrophysical Journal, 952, 44, aDS Bibcode: 2023ApJ...952...44B

Bischetti, M., Piconcelli, E., Vietri, G., et al. 2016, in Active Galactic Nuclei 12: A Multi-Messenger Perspective (AGN12), 12

Borne, K. D. 2009, Astroinformatics: A 21st Century Approach to Astronomy, arXiv:0909.3892 [astro-ph, physics:physics]

Breiman, L. 2001, Machine Learning, 45, 5

Brodzeller, A. & Dawson, K. 2022, The Astronomical Journal, 163, 110

Brownlee, J. 2021, Imbalanced Classification with Python Choose Better Metrics, Balance Skewed Classes, and Apply Cost-Sensitive Learning (Independently published)

Bruni, G., Mack, K. H., Salerno, E., et al. 2012, Astronomy & Astrophysics, 542, A13

Bruni, G., Piconcelli, E., Misawa, T., et al. 2019, Astronomy and Astrophysics, 630, A111, aDS Bibcode: 2019A&A...630A.111B

Buitinck, L., Louppe, G., Blondel, M., et al. 2013, in ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 108–122

Busca, N. & Balland, C. 2018, QuasarNET: Human-level spectral classification and redshifting with Deep Neural Networks, arXiv:1808.09955 [astro-ph]

Canalizo, G. & Stockton, A. 2002, in Astronomical Society of the Pacific Conference Series, Vol. 255, Mass Outflow in Active Galactic Nuclei: New Perspectives, ed. D. M. Crenshaw, S. B. Kraemer, & I. M. George, 195

Capellupo, D. M., Hamann, F., Shields, J. C., Halpern, J. P., & Barlow, T. A. 2013, Monthly Notices of the Royal Astronomical Society, 429, 1872

Capellupo, D. M., Hamann, F., Shields, J. C., Rodríguez Hidalgo, P., & Barlow, T. A. 2011, Monthly Notices of the Royal Astronomical Society, 413, 908

Capellupo, D. M., Hamann, F., Shields, J. C., Rodríguez Hidalgo, P., & Barlow, T. A. 2012, Monthly Notices of the Royal Astronomical Society, 422, 3249

Chamberlain, C., Arav, N., & Benn, C. 2015, Monthly Notices of the Royal Astronomical Society, 450, 1085

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002, Journal of Artificial Intelligence Research, 16, 321

Chen, T. & Guestrin, C. 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (ACM)

Choi, E., Brennan, R., Somerville, R. S., et al. 2020, The Astrophysical Journal, 904, 8

Chollet, F. et al. 2015, Keras, https://github.com/fchollet/keras

Ciesla, L., Charmandaris, V., Georgakakis, A., et al. 2015, Astronomy & Astrophysics, 576, A10

Cuoco, E., Patricelli, B., Iess, A., & Morawski, F. 2021, Universe, 7, 394, aDS Bibcode: 2021Univ....7..394C

De Cicco, D., Bauer, F. E., Paolillo, M., et al. 2021, Astronomy and Astrophysics, 645, A103, aDS Bibcode: 2021A&A...645A.103D

De Cicco, D., Bauer, F. E., Paolillo, M., et al. 2022, Astronomy and Astrophysics, 664, A117, aDS Bibcode: 2022A&A...664A.117D

De Cicco, D., Brandt, W. N., Grier, C. J., & Paolillo, M. 2017, Frontiers in Astronomy and Space Sciences, 4

de Kool, M. 1993, Monthly Notices of the Royal Astronomical Society, 265, L17

de Lima Santos, L. & Soltau, S. B. 2024, The Unified Era: An understanding journey from observations to the Unified Model of Active Galactic Nuclei

DiPompeo, M. A., Runnoe, J. C., Brotherton, M. S., & Myers, A. D. 2013, The Astrophysical Journal, 762, 111

Djorgovski, S. G., Mahabal, A. A., Graham, M. J., Polsterer, K., & Krone-Martins, A. 2022, Applications of AI in Astronomy, arXiv:2212.01493 [astro-ph]

Edge, D. O., Shakeshaft, J. R., McAdam, W. B., Baldwin, J. E., & Archer, S. 1959, Memoirs of the

Royal Astronomical Society, 68, 37

Elvis, M. 2000, The Astrophysical Journal, 545, 63, aDS Bibcode: 2000ApJ...545...63E

Elvis, M., Maccacaro, T., Wilson, A. S., et al. 1978, Monthly Notices of the Royal Astronomical Society, 183, 129

Erakuman, D. & Filiz Ak, N. 2017, Frontiers in Astronomy and Space Sciences, 4

Fabian, A. C. 2012, Annual Review of Astronomy and Astrophysics, 50, 455, arXiv:1204.4114 [astro-ph]

Förster, F., Cabrera-Vives, G., Castillo-Navarrete, E., et al. 2021, The Astronomical Journal, 161, 242, arXiv:2008.03303 [astro-ph]

Fath, E. A. 1909, Lick Observatory Bulletin, 149, 71

Gallagher, S. C., Hines, D. C., Blaylock, M., et al. 2007, The Astrophysical Journal, 665, 157, aDS Bibcode: 2007ApJ...665..157G

Géron, A. 2019, Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow (O'Reilly)

Gaskell, C. M., Gill, J. J. M., & Singh, J. 2016, arXiv e-prints, arXiv:1611.03733

Gautam, S. & Dey, R. 2022, International Research Journal of Computer Science, 9, 89

Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. 2020, Locally Linear Embedding and its Variants: Tutorial and Survey

Gibson, R. R., Brandt, W. N., Schneider, D. P., & Gallagher, S. C. 2008, The Astrophysical Journal, 675, 985, publisher: IOP Publishing

Gibson, R. R., Jiang, L., Brandt, W. N., et al. 2009, The Astrophysical Journal, 692, 758, publisher: The American Astronomical Society

Giustini, M. & Proga, D. 2019, Astronomy & Astrophysics, 630, A94

Green, K. S., Gallagher, S. C., Leighly, K. M., et al. 2023, The Astrophysical Journal, 953, 186, arXiv:2405.06027 [astro-ph]

Green, P. J., Aldcroft, T. L., Mathur, S., Wilkes, B. J., & Elvis, M. 2001, The Astrophysical Journal, 558, 109

Guo, Z. & Martini, P. 2019, The Astrophysical Journal, 879, 72

Halevy, A., Norvig, P., & Pereira, F. 2009, IEEE Intelligent Systems, 24, 8

Hall, P. B., Anderson, S. F., Strauss, M. A., et al. 2002, The Astrophysical Journal Supplement Series, 141, 267

Hamann, F., Chartas, G., McGraw, S., et al. 2013, Monthly Notices of the Royal Astronomical Society, 435, 133

Harrison, C. 2014, PhD thesis, Durham University, UK

Hernández-García, L., Masegosa, J., González-Martín, O., Márquez, I., & Perea, J. 2016, The Astrophysical Journal, 824, 7

Hong, S., Zou, Z., Luo, A. L., et al. 2023, Monthly Notices of the Royal Astronomical Society, 518, 5049

Hopkins, P. F., Torrey, P., Faucher-Giguère, C.-A., Quataert, E., & Murray, N. 2016, Monthly Notices of the Royal Astronomical Society, 458, 816

Hviding, R. E., Hainline, K. N., Goulding, A. D., & Greene, J. E. 2024, The Astronomical Journal, 167, 169, aDS Bibcode: 2024AJ....167..169H

Inoue, A. K., Shimizu, I., Iwata, I., & Tanaka, M. 2014, Monthly Notices of the Royal Astronomical

Society, 442, 1805, aDS Bibcode: 2014MNRAS.442.1805I

Ioffe, S. & Szegedy, C. 2015, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, arXiv:1502.03167 [cs]

Ivezić, Ž., Connolly, A., Vanderplas, J., & Gray, A. 2014, Statistics, Data Mining and Machine Learning in Astronomy (Princeton University Press)

Iwasaki, D., Cooray, S., & Takeuchi, T. T. 2023, Extracting an Informative Latent Representation of High-Dimensional Galaxy Spectra, publication Title: arXiv e-prints ADS Bibcode: 2023arXiv231117414I

Jolliffe, I. T. & Cadima, J. 2016, Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 374, 20150202

Kan Ho, T. 2016, Random Decision Forests

Kao, W.-B., Zhang, Y., & Wu, X.-B. 2024, Efficient Identification of Broad Absorption Line Quasars using Dimensionality Reduction and Machine Learning, arXiv:2404.12270 [astro-ph]

Kasliwal, V. P., Vogeley, M. S., & Richards, G. T. 2015, Monthly Notices of the Royal Astronomical Society, 451, 4328

Kingma, D. P. & Ba, J. 2017, Adam: A Method for Stochastic Optimization, arXiv:1412.6980 [cs]

Kolmogorov-Smirnov, A., Kolmogorov, A. N., & Kolmogorov, M. 1933

Kovacevic, A., Ilic, D., Jankov, I., et al. 2021, LSST AGN SC Cadence Note: Two metrics on AGN variability observable

Krawczyk, C. M., Richards, G. T., Gallagher, S. C., et al. 2015, The Astronomical Journal, 149, 203

Krawczyk, C. M., Richards, G. T., Mehta, S. S., et al. 2013, The Astrophysical Journal Supplement Series, 206, 4, aDS Bibcode: 2013ApJS..206....4K

Kumar, P., Bhatnagar, R., Gaur, K., & Bhatnagar, A. 2021, IOP Conference Series: Materials Science and Engineering, 1099, 012077

Labach, A., Salehinejad, H., & Valaee, S. 2019, Survey of Dropout Methods for Deep Neural Networks, arXiv:1904.13310 [cs]

Lawrence, A., Warren, S. J., Almaini, O., et al. 2007, Monthly Notices of the Royal Astronomical Society, 379, 1599

Leighly, K. M., Choi, H., Eracleous, M., et al. 2024, The Astrophysical Journal, 966, 87

Lemaître, G., Nogueira, F., & Aridas, C. K. 2017, Journal of Machine Learning Research, 18, 1

Levene, H. 1961

Lewis, G. F., Chapman, S. C., & Kuncic, Z. 2003, The Astrophysical Journal Letters, 596, L35

Liang, P. P., Zadeh, A., & Morency, L.-P. 2023, Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions, arXiv:2209.03430 [cs]

Lira, P., Arévalo, P., Uttley, P., McHardy, I. M. M., & Videla, L. 2015, Monthly Notices of the Royal Astronomical Society, 454, 368, aDS Bibcode: 2015MNRAS.454..368L

Liu, K., Li, Y., Xu, N., & Natarajan, P. 2018, Learn to Combine Modalities in Multimodal Deep Learning, arXiv:1805.11730 [cs, stat]

Liu, Y., Jin, J., Zhao, H., He, X., & Guo, Y. 2023, The Astrophysical Journal, 954, 86

LSST-Science-Collaboration, Abell, P. A., Allison, J., et al. 2009, LSST Science Book, Version 2.0, arXiv:0912.0201 [astro-ph]

Lundgren, B. F., Wilhite, B. C., Brunner, R. J., et al. 2007, The Astrophysical Journal, 656, 73

Lyke, B. W., Higley, A. N., McLane, J. N., et al. 2020, The Astrophysical Journal Supplement Series, 250, 8, arXiv:2007.09001 [astro-ph]

Lynds, C. R. 1967, The Astrophysical Journal, 147, 396, aDS Bibcode: 1967ApJ...147..396L

Lynds, R. 1971, The Astrophysical Journal, 164, L73

Maddox, N. & Hewett, P. C. 2008, Memorie della Società Astronomica Italiana, 79, 1117

Malanchev, K. L., Pruzhinskaya, M. V., Korolev, V. S., et al. 2021, MNRAS, 502, 5147

Möller, A., Peloton, J., Ishida, E. E. O., et al. 2020, Monthly Notices of the Royal Astronomical Society, 501, 3272

Marshall, A., Auger-Williams, M. W., Banerji, M., Maiolino, R., & Bowler, R. 2022, Monthly Notices of the Royal Astronomical Society, 515, 5617–5628

Mas-Ribas, L. & Mauland, R. 2019, The Astrophysical Journal, 886, 151

Masci, F. J., Laher, R. R., Rusholme, B., et al. 2018, Publications of the Astronomical Society of the Pacific, 131, 018003

Matheson, T., Stubens, C., Wolf, N., et al. 2021, The Astronomical Journal, 161, 107

McGraw, S. M., Shields, J. C., Hamann, F. W., Capellupo, D. M., & Herbst, H. 2017, Monthly Notices of the Royal Astronomical Society, 475, 585

McInnes, L., Healy, J., & Melville, J. 2020, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv:1802.03426 [cs, stat]

Menou, K., Vanden Berk, D. E., Ivezić, Ž., et al. 2001, The Astrophysical Journal, 561, 645

Miller, T. R., Arav, N., Xu, X., & Kriss, G. A. 2020, Monthly Notices of the Royal Astronomical Society, 499, 1522

Mishra-Sharma, S., Song, Y., & Thaler, J. 2024, PAPERCLIP: Associating Astronomical Observations and Natural Language with Multi-Modal Models

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. 2019, M3ER: Multiplicative Multimodal Emotion Recognition Using Facial, Textual, and Speech Cues

Montenegro-Montes, F. M., Mack, K. H., Benn, C. R., et al. 2009, arXiv e-prints, arXiv:0903.5119

Mountrichas, G., Buat, V., Yang, G., et al. 2021, Astronomy & Astrophysics, 646, A29

Naddaf, M. H., Martinez-Aldama, M. L., Marziani, P., et al. 2023, Astronomy and Astrophysics, 675, A43, aDS Bibcode: 2023A&A...675A..43N

Nair, A. & Vivek, M. 2022, Monthly Notices of the Royal Astronomical Society, 511, 4946

Netzer, H. 2015, Annual Review of Astronomy and Astrophysics, 53, 365

Ngiam, J., Khosla, A., Kim, M., et al. 2001

Nordin, J., Brinnel, V., van Santen, J., et al. 2019, Astronomy and Astrophysics, 631, A147

Odewahn, S. C. 1998, in American Astronomical Society Meeting Abstracts, Vol. 192, American Astronomical Society Meeting Abstracts #192, 64.04

Otsu, N. 1979, IEEE Transactions on Systems, Man, and Cybernetics, 9, 62

Parcalabescu, L., Trost, N., & Frank, A. 2021, What is Multimodality?, arXiv:2103.06304 [cs]

Parker, L., Lanusse, F., Golkar, S., et al. 2024, Monthly Notices of the Royal Astronomical Society, 531, 4990

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research,

12, 2825

Petley, J. W., Morabito, L. K., Alexander, D. M., et al. 2022, Monthly Notices of the Royal Astronomical Society, 515, 5159, publisher: OUP ADS Bibcode: 2022MNRAS.515.5159P

Portillo, S. K. N., Parejko, J. K., Vergara, J. R., & Connolly, A. J. 2020, The Astronomical Journal, 160, 45, publisher: The American Astronomical Society

Probst, P., Wright, M., & Boulesteix, A.-L. 2019, WIREs Data Mining and Knowledge Discovery, 9, arXiv:1804.03515 [cs, stat]

Pruzhinskaya, M. V., Ishida, E. E. O., Novinskaya, A. K., et al. 2023, Astronomy &; Astrophysics, 672, A111

Ramos Almeida, C. & Ricci, C. 2017, Nature Astronomy, 1, 679

Rankine, A. L., Hewett, P. C., Banerji, M., & Richards, G. T. 2020, Monthly Notices of the Royal Astronomical Society, 492, 4553

Reichard, T. A., Richards, G. T., Hall, P. B., et al. 2003, The Astronomical Journal, 126, 2594

Ricci, C. & Trakhtenbrot, B. 2023, Nature Astronomy, 7, 1282

Richards, G. T., Lacy, M., Storrie-Lombardi, L. J., et al. 2006, The Astrophysical Journal Supplement Series, 166, 470

Robinson, L., Grier, K., Horne, K., et al. 2024, 56, 105.01, conference Name: American Astronomical Society Meeting Abstracts ADS Bibcode: 2024AAS...24410501R

Rodríguez Hidalgo, P. & Rankine, A. L. 2022, The Astrophysical Journal Letters, 939, L24

Ruan, J. J., Anderson, S. F., Green, P. J., et al. 2016, The Astrophysical Journal, 825, 137

Saccheo, I., Bongiorno, A., Piconcelli, E., et al. 2023, Astronomy and Astrophysics, 671, A34, aDS Bibcode: 2023A&A...671A..34S

Salpeter, E. E. 1964, The Astrophysical Journal, 140, 796

Sánchez, P., Lira, P., Cartier, R., et al. 2017, The Astrophysical Journal, 849, 110, publisher: The American Astronomical Society

Sánchez-Sáez, P., Lira, H., Martí, L., et al. 2021a, The Astronomical Journal, 162, 206, arXiv:2106.07660 [astro-ph]

Sánchez-Sáez, P., Lira, P., Mejía-Restrepo, J., et al. 2018, The Astrophysical Journal, 864, 87

Sánchez-Sáez, P., Reyes, I., Valenzuela, C., et al. 2021b, The Astronomical Journal, 161, 141, aDS Bibcode: 2021AJ....161..141S

Saraswat, P. & Jain, D. 2021, International Journal of Advanced Engineering Research and Applications, 7, 40

Savić, D. V., Jankov, I., Yu, W., et al. 2023, The LSST AGN Data Challenge: Selection methods, publication Title: arXiv e-prints ADS Bibcode: 2023arXiv230704072S

Schmidt, M. 1963, Nature Astronomy, 197, 1040

Seyfert, C. K. 1943, The Astrophysical Journal, 97, 28

Shen, Y., Richards, G. T., Strauss, M. A., et al. 2011, The Astrophysical Journal Supplement Series, 194, 45, aDS Bibcode: 2011ApJS..194...45S

Sheng, X., Ross, N., & Nicholl, M. 2022, Monthly Notices of the Royal Astronomical Society, 512, 5580–5600

Shields, G. A. 1999, Publications of the Astronomical Society of the Pacific, 111, 661

Shukla, S. N. & Marlin, B. M. 2021, Multi-Time Attention Networks for Irregularly Sampled Time

Series

Shwartz-Ziv, R. & Armon, A. 2021, Tabular Data: Deep Learning is Not All You Need, arXiv:2106.03253 [cs]

Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, The Astronomical Journal, 131, 1163

Sleeman IV, W. C., Kapoor, R., & Ghosh, P. 2021, Multimodal Classification: Current Landscape, Taxonomy and Future Directions, arXiv:2109.09020 [cs]

Smith, M. J. & Geach, J. E. 2023, Royal Society Open Science, 10, 221454, aDS Bibcode: 2023RSOS...1021454S

Sniegowska, M., Naddaf, M. H., Martinez-Aldama, M. L., et al. 2023, arXiv e-prints, arXiv:2306.03224

Sokol, A. D., Yun, M., Pope, A., Kirkpatrick, A., & Cooke, K. 2023, Monthly Notices of the Royal Astronomical Society, 521, 818

Son, S., Kim, M., & Ho, L. C. 2023, The Astrophysical Journal, 958, 135, aDS Bibcode: 2023ApJ...958..135S

Spinoglio, L. & Fernández-Ontiveros, J. A. 2019, Proceedings of the International Astronomical Union, 15, 29, arXiv:1911.12176 [astro-ph]

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. 2014, Journal of Machine Learning Research, 15, 1929

Szakacs, R., Péroux, C., Nelson, D., et al. 2023, The BarYon CYCLE Project (ByCycle): Identifying and Localizing MgII Metal Absorbers with Machine Learning, publication Title: arXiv e-prints ADS Bibcode: 2023arXiv230517970S

Tammour, A., Gallagher, S. C., Daley, M., & Richards, G. T. 2016, Monthly Notices of the Royal Astronomical Society, 459, 1659

Teimoorinia, H., Archinuk, F., Woo, J., Shishehchi, S., & Bluck, A. F. L. 2022, The Astronomical Journal, 163, 71, publisher: The American Astronomical Society

Temple, M. J., Hewett, P. C., & Banerji, M. 2021, Monthly Notices of the Royal Astronomical Society, 508, 737, aDS Bibcode: 2021MNRAS.508..737T

Torres-Papaqui, J. P., Coziol, R., Robleto-Orus, A. C., Cutiva-Alvarez, K. A., & Roco-Avilez, P. 2024, The role of AGN winds in galaxy formation: connecting AGN outflows at low redshifts to the formation/evolution of their host galaxies, arXiv:2405.05184 [astro-ph]

Trump, J. R., Hall, P. B., Reichard, T. A., et al. 2006, The Astrophysical Journal Supplement Series, 165, 1

Ulrich, M.-H., Maraschi, L., & Urry, C. M. 1997, Annual Review of Astronomy and Astrophysics, 35, 445

Urry, C. M. & Padovani, P. 1995, Publications of the Astronomical Society of the Pacific, 107, 803

van der Maaten, L. & Hinton, G. 2008, Journal of Machine Learning Research, 9, 2579

van Dyk, D. A. & Meng, X.-L. 2001, Journal of Computational and Graphical Statistics, 10, 1

Vanden Berk, D. E., Richards, G. T., Bauer, A., et al. 2001, The Astronomical Journal, 122, 549, aDS Bibcode: 2001AJ....122..549V

Vanderplas, J., Connolly, A., Ivezić, Ž., & Gray, A. 2012, in Conference on Intelligent Data Understanding (CIDU), 47 –54

Vaswani, A., Shazeer, N., Parmar, N., et al. 2023, Attention Is All You Need

Wasserstein, R. L. & Lazar, N. A. 2016, The American Statistician, 70, 129

Webb, S. A. & Goode, S. R. 2023, An Astronomers Guide to Machine Learning, publication Title: arXiv e-prints ADS Bibcode: 2023arXiv230400512W

Welling, C. A., Miller, B. P., Brandt, W. N., Capellupo, D. M., & Gibson, R. R. 2014, Monthly Notices of the Royal Astronomical Society, 440, 2474

Weymann, R. J., Morris, S. L., Foltz, C. B., & Hewett, P. C. 1991, The Astrophysical Journal, 373, 23, aDS Bibcode: 1991ApJ...373...23W

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, The Astronomical Journal, 140, 1868, aDS Bibcode: 2010AJ....140.1868W

Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., & Le, Q. V. 2020, Unsupervised Data Augmentation for Consistency Training

Zhang, S., Wang, H., Wang, T., et al. 2014, The Astrophysical Journal, 786, 42

Zhao, F., Zhang, C., & Geng, B. 2024, ACM Comput. Surv., 56

Zhao, S., Song, J., & Ermon, S. 2017, ArXiv

Zubovas, K. 2018, Monthly Notices of the Royal Astronomical Society, 479, 3189

Acknowledgments

With this piece of work, I somehow reach the end of this journey, still existing as (not quite) the same person as day one. The craft of science, ever so infinite, gasp-inducing, undeniably overwhelming and terrifying too, and deeply complex, is, beware, only for the outliers, the nerds and the deeply passionate ones. And yet, the only ever appropriate approach is team-work. There is no such thing as a successful isolated scientist. To all the colleagues who made this work possible, and who have supported me in this path, thank you:

To Angela Bongiorno and Andjelka Kovačević, for guiding me through the tumultuous waters of my first thesis with great kindness and patience, and for Angela's "done is better than perfect" motto.

To Ivano Saccheo, for finding all my typos and helping me with all my random doubts.

To Francesco Tombesi and Enrico Piconcelli, for your helpful feedback and ideas.

To Paola Marziani, for sharing the sample that acted as main character throughout this work.

To Anais Möller, Francisco Förster and Lars Doorenbos, for their insightful questions and ideas during the ML4Astro conference in Catania, which greatly contributed to the machine learning work in this thesis.

To Manuel Merello and Leo Bronfman, for kickstarting my astronomical path with such enthusiasm.

To Paula Sánchez Sáez, for taking me under your wing.

And finally, to my past self. For persevering.

You can go home again, the General Temporal Theory asserts, so long as you understand that home is a place where you have never been.
- Ursula K. Le Guin, The Dispossessed: An Ambiguous Utopia

Nicolás G. Guerra Varas acknowledges support through an Erasmus Mundus Joint Master (EMJM) scholarship funded by the European Union in the framework of the Erasmus+, Erasmus Mundus Joint Master in Astrophysics and Space Science – MASS. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or granting authority European Ed- ucation and Culture Executive Agency (EACEA). Neither the European Union nor the granting authority can be held responsible for them.