Universidad de los Andes

Faculty of Sciences

Mapping the Universe: classifying galaxies and quasars through essential spectral features with machine learning

Thesis presented by Daniel Fajardo

to obtain the degree of Physics

Advisor Dr. Alejandro García

June 2024

Physics department

"The poets leave hell and again behold the stars."

-Dante Alighieri.

Acknowledgments

I would like to express my sincere gratitude to professor Alejandro García for his invaluable support and guidance throughout the development of this thesis. His experience and commitment have been crucial in shaping the ideas and conducting meaningful research. Additionally, I thank Manuel Alejandro Segura Delgado for their outstanding procedural, statistical, and computational support. I am grateful for the opportunity to work alongside such talented and dedicated professionals.

I would also like to express my gratitude to my parents and colleagues, Carlos Fajardo, Martha Poveda, Johan Fajardo Poveda, Camila Cárdenas Uribe, and Julián Rojas Tapias, for their support, friendship and patience. I thank them for understanding me in all my academic and personal stages. Without their company, everything would have been more difficult.

Abstract

Astronomers today have access to instruments capable of making extremely precise measurements of the spectra and light curves of objects such as stars, quasars, and galaxies. The exciting part is when the data acquisition rate is extremely efficient, granting us access to information on millions of objects. However, this has driven the scientific community to harness all available computational power to process these vast amounts of data more efficiently. There are some classifiers like RedRock or QuasarNET that perform spectral classification and redshift measurements for the targets observed by DESI. However, not all results obtained for all types of targets are satisfactory; there are ambiguous boundaries between the redshifts of galaxies and QSOs, and negative values for the redshift of very distanced objects that are poorly estimated. This complicates and biases the analysis of experimental data.

In the pursuit of enhancing classification capabilities, powerful tools like H2O have emerged. These resources transcend traditional approaches, conducting comprehensive training on our data, optimizing models, and suggesting effective machine learning methods for analyzing our spectra and input parameters. Our study endeavors to compare the outcomes produced by different models, such as Random Forest, LightGBM, with those of the neural network SAM, designed by us.

The objective is to assess and contrast the effectiveness of each model in accurately classifying spectral types by using spectrophotometric features. In the end, we are going to handle with the polarimetry problem of observing quasars in different angles, and we try to expose the problem of using just spectral lines to distinguish between quasars and galaxies.

Contents

Chapter 1

Introduction

1.1 Relevance of emission lines in astronomical spectra

In 1880, Kirchhoff and Bunsen developed the idea that every element produces a proper and unique pattern of emission and absortion spectral lines. Kirchhoff tried to reduce all the phenomena of spectral lines production in three laws, now called the **Kirchhoff's laws** (Ostlie & Carroll 1996):

1. A luminous solid, liquid, or dense gas emits light of all wavelengths.

2. A low density, hot gas seen against a cooler background emits a bright line or emission line spectrum.

3. A low density, cool gas in front of a hotter source of a continuous spectrum creates a dark line or absorption line spectrum.

These laws underpin the understanding of spectra, where electromagnetic radiation is dispersed based on its wavelength. Spectra are commonly represented as plots of flux (intensity or brightness) against wavelength, providing a visual depiction of an object's unique spectral fingerprint.

The spectra emitted by celestial bodies serve as invaluable tools for understanding their

composition. Each element and molecule absorbs and emits light at distinct wavelengths, creating characteristic lines in the spectrum known as emission and absorption lines, as shown in Figure 1.1. By analyzing these lines, astronomers can identify the elements present in a body. For instance, hydrogen often exhibits prominent Balmer lines, while heavier elements contribute their own distinctive patterns (Karttunen et al. 2007). Absorption lines occur when cooler outer layers of a celestial object absorb specific wavelengths, leaving dark lines in the spectrum. These spectral features offer a window into the chemical makeup of stars, galaxies, quasars, and other astronomical entities.

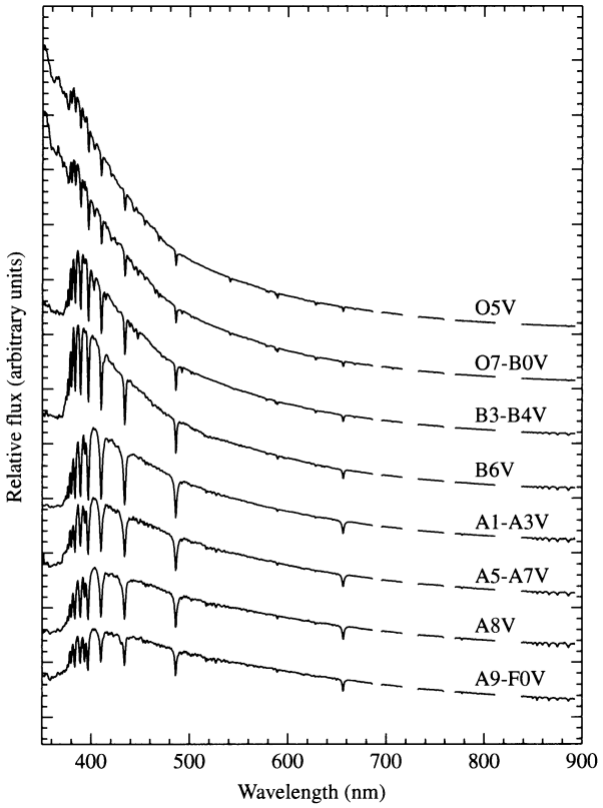

Figure 1.1: Spectra of main sequence stars displayed in terms of the flux as a function of the wavelenght. Extracted from (Ostlie & Carroll 1996).

The Doppler effect, a fundamental concept in spectroscopy, manifests as a shift in the observed wavelength of light due to relative motion between the source and the observer. This phenomenon is evident in both emission and absorption lines. When an object is moving away, the lines shift toward the longer wavelength (redshift), indicating recession.

Conversely, a blueshift occurs when the object approaches, leading to shorter wavelength lines. Astronomers exploit the redshift to estimate the radial velocity of celestial objects and, consequently, their distance. Redshift values provide crucial information about the universe's expansion and the vast distances between galaxies. Some of the most useful relations are given by the expressions $z = \frac{\lambda_{obs} - \lambda_{rest}}{\lambda_{rest}} = \frac{v_r}{c}$, and $z = \frac{H_0}{c} d$. Here, z is the redshift index, v_r the radial velocity, c the speed of the light, H_0 the Hubble constant, and d the distance of the observed object.

The spectrum, a treasure trove of information, extends beyond identifying elemental composition and motion. Astronomers associate spectra with various properties, including the effective temperature of stars, their chemical composition, and even their age. Moreover, the redshift obtained from a spectrum enables astronomers to determine a celestial object's distance and, consequently, its position in the vast cosmic tapestry. Spectroscopy emerges as an indispensable tool, allowing scientists to unravel the mysteries of the universe and glean insights into the nature of distant galaxies, stars, and cosmic phenomena.

1.2 DESI project

The Dark Energy Spectroscopic Instrument (DESI) is an astronomical initiative with the primary goal of advancing our understanding of dark energy, a mysterious force thought to be responsible for the Universe's accelerating expansion. Installed in the 4-meter Mayall telescope at the Kitt Peak National Observatory in Arizona, USA, DESI is a sophisticated astronomical spectrograph. Its primary mission involves conducting an ambitious program of spectroscopic observations on distant galaxies across a wide swath of the celestial sphere (Adame et al. 2023).

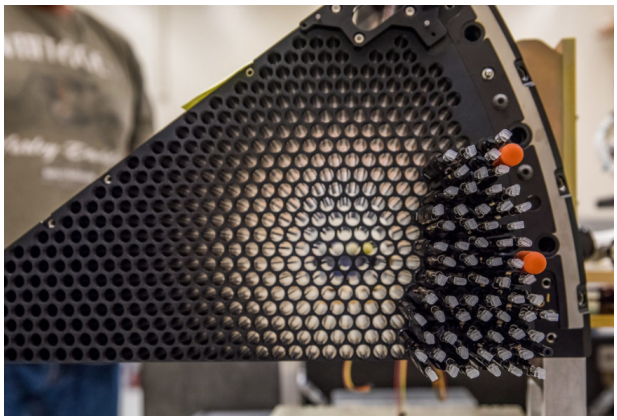

Figure 1.2: One of the 10 petal-shaped focal planes of the DESI spectrograph, equipped with 500 optical fibers capable of receiving and focusing light from the targets. Extracted from DESI Collaboration website.

To achieve this, DESI employs a network of 5020 optical fibers, of which 20 are designated for monitoring brightness through a camera, while the remaining 5000 are pointed toward specific celestial targets. These fibers guide light to one of the 10 spectrographs, refers to Figure 1.2, which collect data regarding the wavelengths of light emitted by billions of galaxies and quasars throughout the Universe. The focal plane is composed of 10 distinct petals, each corresponding to a spectrograph that can simultaneously read data from 500 fibers[1]. Each spectrograph is further equipped with three cameras, each of which is dedicated to a specific wavelength filter: B (3600-5800) Å, R (5760-7620) Å, Z (7520-9824) Å (Adame et al. 2023). It is expected that by using the data gathered by DESI, we can gain insights into the nature of dark energy, dark matter, galactic evolution, distant supernovae, and cosmic structure mapping.

1.3 The Cannon and some useful Python modules

One of the earliest applications of artificial intelligence in the field of astronomy was showcased by *The Cannon*. This technology demonstrated the ability to model the dependence of

[1]You can read more about the focal plane system in the DESI collaboration page: https://www.desi.lbl.gov/focal-plane-system/

normalized continuum spectra on specific characteristic features, such as metallicity ([Fe/H]), logarithm of gravity (Ln(g)), and effective temperature (Te). The training dataset for this artificial intelligence comprised primarily 542 stars from 19 different clusters. Since these stars generally share similar characteristics, the training process is facilitated (Ness et al. 2015).

Subsequently, the goal was to assess the model's performance by validating it against the spectra of 55,000 stars from APOGEE DR10. It is important to note that this model does not integrate a rigorous physical analysis; instead, each step of the process, including the identification of continuum regions and linear or non-linear fitting, is implemented manually, as seen in Figure 1.3. Despite this manual intervention, the model produced satisfactory results, exhibiting only a 30% increase in error compared to APOGEE DR10 (Ness et al. 2015).

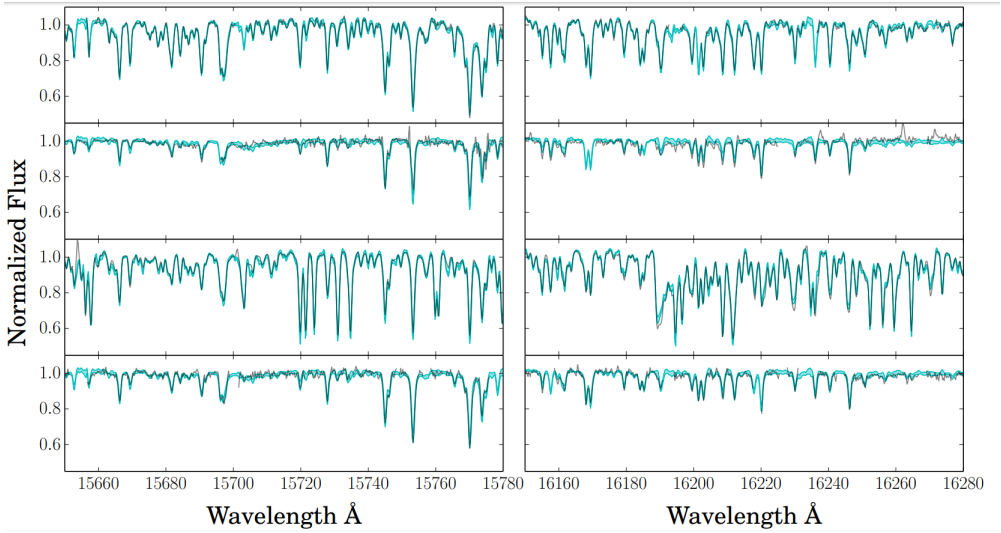

Figure 1.3: Spectral fitting with The Cannon's generative model. The original flux data is depicted in black, while the fitted data is shown in blue. Extracted from (Ness et al. 2015).

Nowadays, specialized libraries in languages like Python offer a more convenient approach to developing tools akin to *The Cannon*, eliminating the need for manual implementation at each step. Notable among these libraries are *Pyspeckit*, *Keras*, *TensorFlow*, and *Astropy*.

Within *Pyspeckit*, a spectrum is represented as an object with three key attributes:

the Plotter, the fitter specfit, and the continuum fitter baseline. These attributes prove invaluable for the selection of spectral regions to be fitted. The fitting process itself utilizes the Levenberg–Marquardt optimization method facilitated by the *mpfit* and *lmfit* libraries (Adam 2022). For error analysis, *Pyspeckit* leverages *pymc* and *emcee*. Additionally, *Pyspeckit* offers an array of base models, such as Gaussian, Lorentzian, Voigt profiles, as well as models tailored for specific applications like ammonia, hyperfine formaldehyde, H2 rotational ladder, and recombination lines (Ginsburg et al. 2022).

In *Astropy*, a spectrum is conceptualized, also, as an object encapsulating several essential attributes. These attributes play a crucial role in facilitating the selection of spectral regions for fitting purposes. The fitting process in *Astropy* makes use of optimization methods, with functionalities drawn from its extensive modeling capabilities. Unlike *Pyspeckit*, *Astropy* offers a broader scope, extending beyond spectra to encompass diverse data reading formats (Robitaille et al. 2013). The library supports various applications and includes features such as spectral data loading, visualization, model fitting, property measurement, line corrections, data processing, interactive interaction, and results export. Similar to *Pyspeckit*, *Astropy* provides a versatile platform for astronomers and researchers, allowing for efficient manipulation and analysis of astronomical data.

In *Keras* and *TensorFlow*, these powerful machine learning frameworks provide a comprehensive set of tools for building, training, and deploying various types of machine learning models. *Keras* serves as a high-level neural networks API, offering an intuitive interface for constructing and experimenting with deep learning architectures. It emphasizes ease of use and user-friendly syntax, making it particularly accessible for both beginners and experienced practitioners.

TensorFlow, on the other hand, is an open-source machine learning library developed by the Google Brain team. It provides a flexible platform for constructing and training machine learning models, with a focus on scalability and deployment across various devices and platforms. *Keras* is now tightly integrated into *TensorFlow* as its official high-level API,

offering a streamlined and unified approach to building deep learning models.

Both *Keras* and *TensorFlow* are designed to handle a wide range of machine learning tasks, not limited to deep learning. They support various types of models, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and more, making them suitable for diverse applications such as image recognition, natural language processing, and regression tasks.

One of the notable strengths of these frameworks is their compatibility with powerful hardware accelerators, allowing for efficient training and inference on GPUs and TPUs. Additionally, their extensive documentation, vibrant community support, and integration with other popular libraries make them preferred choices for researchers and developers in the machine learning and artificial intelligence domains.

1.4 Spectrum classifiers and redshift estimators

The DESI project has implemented various tools, such as fast classifiers like *RedRock* and Convolutional Neural Networks (CNNs) like *QuasarNET* and *SQUEEzE*, to predict the spectral type and redshift of the measured targets. Although the operation of these classifiers differs significantly, they have proven to be highly valuable for spectrum analysis. Additionally, in the analysis of catalogs generated through different strategies, where the outputs of both classifiers were collaboratively integrated, it was determined that the $QN|RR$ strategy, involving the acceptance of any object cataloged as a quasar by at least one of the two classifiers, is the most effective. This strategy achieves the optimal balance between catalog contamination and catalog completeness (Farr et al. 2020).

1.4.1 RedRock

This package utilizes a set of spectral type templates to evaluate the resemblance between an input spectrum and these templates, generating predictions accordingly (Adame et al. 2023).

The methodology involves specifying a particular redshift range for each spectral type; for instance, quasars were assigned a range of $0.0033 < Z < 7$, look at Figure 1.4. The input spectrum is then adjusted along the wavelength axis for each Z value using the Principal Component Analysis (PCA) method, representing the spectrum as a linear combination of the principal components of the templates (Saavedra 2019).

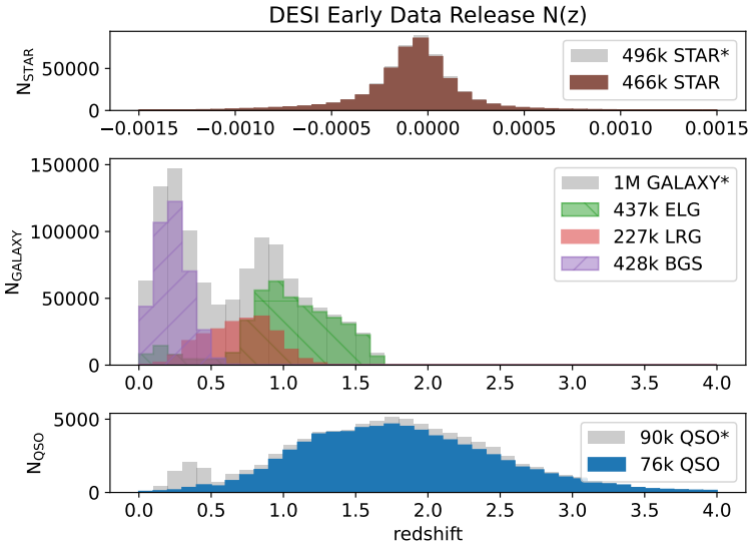

Figure 1.4: Redshift ranges and distributions per spectral type. Extracted from (Adame et al. 2023).

Subsequently, a χ^2 value is computed for each fit through the homogeneity test, and the fit with the lowest χ^2 value is considered the best fit. This process allows for estimating the redshift and spectral type using Redrock. Additionally, this method provides the uncertainty of the redshift value (ZERR) and a warning flag (ZWARN) that identifies potential issues or biases in the prediction.

While demonstrating commendable performance in minimizing contamination at redshifts equal to or greater than 2.2, the method faced significant challenges at lower redshifts. Notably, it introduced substantial contamination from stellar sources and exhibited an omission rate of over 5% for Quasi-Stellar Objects (QSOs) across all redshifts. Consequently, the accuracy of this approach fell short of meeting the criteria deemed necessary for conclusive classifications. DESI is actively working on the development of new Quasi-Stellar Object (QSO) templates.

The primary focus of this endeavor is to enhance the performance of the existing method, particularly for QSOs at redshifts below 2.2. It is worth noting that while these newly crafted Redrock templates are in the developmental phase, they have not yet been made available for implementation.

1.4.2 QuasarNET and SQUEzE

In addition to employing fast classifiers like Redrock, convolutional neural networks (CNNs) were utilized to enhance the accuracy of spectral classification and redshift measurement. QuasarNET, designed to address cases of ambiguous classifications between galaxies and quasars produced by algorithms like Redrock (Farr et al. 2020), is one such CNN. Another relevant CNN, SQUEzE, shares similarities with QuasarNET, and both aim to rectify previous results. These CNNs identify significant emission lines and sequences within emission peaks, employing distinct methodologies. SQUEzE focuses on a defined flux range close to 1, while QuasarNET targets a specific range of redshift values, particularly for QSOs with values exceeding $Z > 2.1$. Additionally, QuasarNET incorporates information from the reobservation of these objects and utilizes the predictions made by Redrock as an initial reference, a strategy known as $QN|RR$ (Saavedra 2019).

QuasarNET attempts to emulate human identification of emission lines by using the smoothed spectrum as input. The entry passes through four convolutional layers, where the CNN identifies the main patterns. The output from these convolutional layers is then directed to a fifth dense or fully connected layer before feeding into several "line finder" units, which ultimately identify specific emission lines. Specifically, QuasarNET is trained to detect the $Ly\alpha$, $C[IV]$, $C[III]$, $Mg[II]$, $H\beta$, and $H\alpha$ lines.

In contrast, SQUEzE takes a markedly distinct approach compared to QuasarNET. The SQUEzE methodology begins by smoothing each spectrum to eliminate noise features and subsequently searches for emission peaks surpassing a predetermined significance threshold. Spectra without significant peaks are discarded, while those with noteworthy peaks are

retained. These retained spectra, treated as features, are then input into a random forest classifier. SQUEzE strategically restricts the features seen by the random forest to high-level metrics, aiming to mitigate learning from spurious features in the training set, such as instrumental defects or pipeline reduction errors. The random forest evaluates the validity of each trial peak-identity pair, assigning a "confidence" score to indicate the likelihood of a correct identification. Ultimately, the spectrum receives a QSO classification if the highest confidence score surpasses a specified threshold, a value tailored to the specific requirements of the classification task at hand.

Finally, a QSO catalog constructed through simultaneous predictions by QuasarNET and Redrock ($QN|RR$) significantly reduces contamination by a factor of at least 2, improving accuracy from 99% to 99.5% in QSOs with high redshift, albeit with some completeness reduction. This results are shown in Figure 1.5.

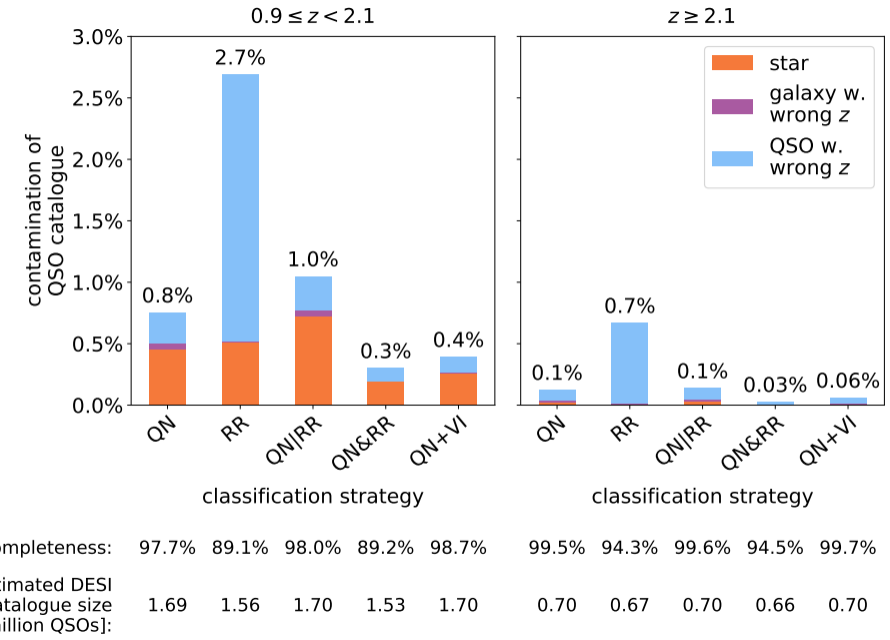

Figure 1.5: Contamination and Completeness of Catalogs Generated Using Different Classification Strategies. In this context, QN refers to the catalog of quasars classified by QuasarNET, while RR denotes the catalog created by the RedRock classifier. Additionally, the notation $QN\&RR$ represents the strategy of simultaneous classification by both classifiers, whereas $QN|RR$ signifies a strategy where an object is treated as a quasar if classified as such by either of the two classifiers. Extracted from (Farr et al. 2020).

1.5 The Quasars's nature

Quasars are fascinating astronomical objects that represent one of the most energetic and extreme phenomena in the universe. They are active galactic nuclei powered by supermassive black holes in the process of accreting matter. These black holes, with masses that can reach millions or even billions of times the mass of the Sun, are surrounded by accretion disks, formed by gas and dust orbiting around them at extremely high speeds, about 1000's $km\ s^{-1}$ (Gallagher et al. 2015).

The continuous optical emission through the ultraviolet in quasars is generated by this optically thick material in the accretion disk, located at the center of the host galaxy. This disk is fundamentally important as it is the primary source of energy for quasars. At extremely close distances to the black hole, of the order of a few gravitational radii (Gallagher et al. 2015), this material heats up to extremely high temperatures due to friction and the intense gravity of the black hole, emitting radiation in the form of visible and ultraviolet light (4,000-10,000) Å.

One of the most intriguing aspects of quasars is the presence of broad emission lines in their optical spectra, known as Seyfert type I lines (Zakamska & Alexandroff 2023). These broad lines are indicative of the presence of highly ionized and fast-moving gas around the central black hole. In many cases, these broad emission lines show signs of winds, which are streams of gas and dust expelled at high speeds from the active galactic nucleus. These winds, driven by the intense radiation emitted by the central black hole, can have a significant impact on the evolution of the host galaxy.

Espectropolarimetry, a technique used in astronomy to study the polarization of light emitted by astronomical objects, has been instrumental in the study of quasars and their distinction from other astronomical objects. This technique provides valuable information about the geometry and kinematics of the objects studied, as well as the properties of the interstellar medium surrounding them, which helps better understand the nature of quasars

and their role in the universe (Zakamska & Alexandroff 2023). In some quasars, the light emitted by the active nucleus can be absorbed by a torus of gas and dust surrounding the accretion disk and then reemitted in the form of narrow spectral lines. This phenomenon, known as narrow-line emission, is characteristic of Seyfert type II quasars. These narrow lines originate in the ionized gas within the torus and are the result of absorption and reemission processes of radiation. The Figure 1.6 displays examples of quasar spectra, showcasing both the broad and narrow emission lines described earlier.

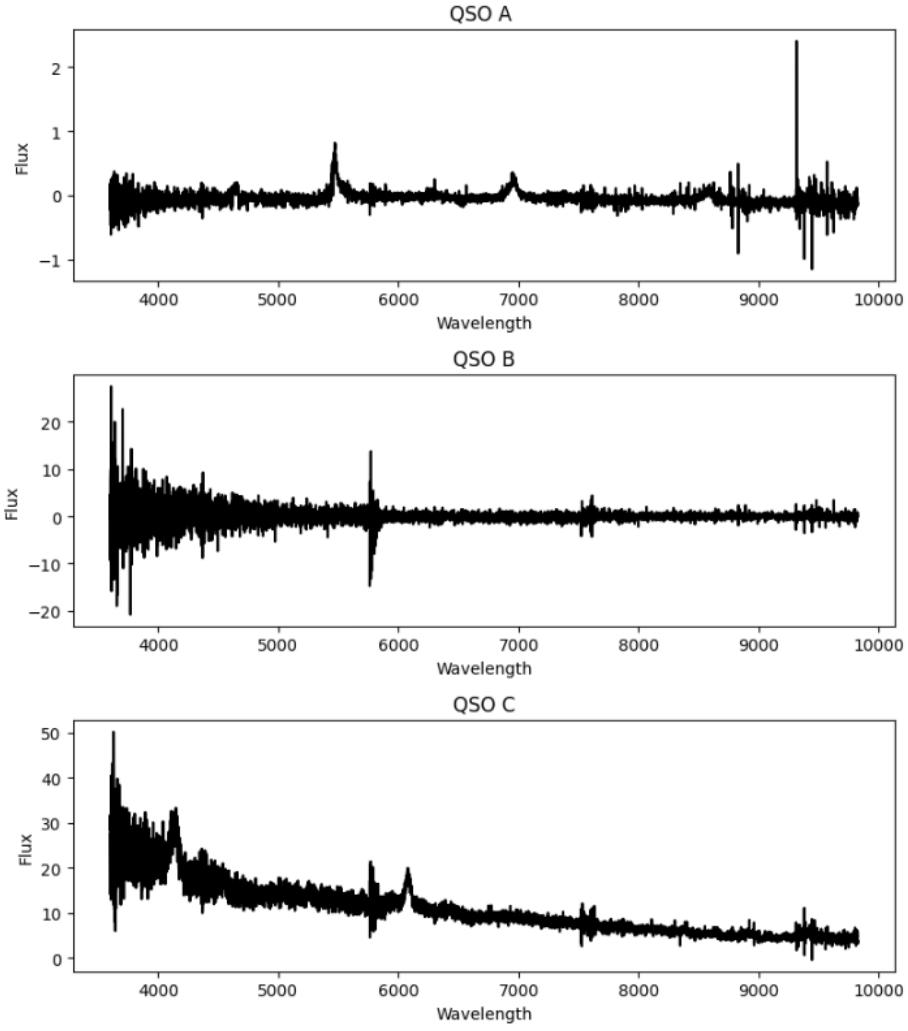

Figure 1.6: Quasar Spectra. Spectrum A displays a combination of broad and narrow lines. Conversely, spectrum B potentially depicts a quasar observed directly through its torus, as evidenced by the absence of broad lines and predominance of narrow ones. Lastly, spectrum C exhibits a substantial presence of broad lines alongside few narrow lines, suggesting potential direct observation of the active nucleus.

The presence of broad emission lines in the optical spectra of quasars can make it difficult to distinguish them from other astronomical objects, such as stars and galaxies. This is because Seyfert galaxies, a specific type of galaxies with active nuclei, can also present broad and narrow emission lines in their optical spectra, this can be observed in Figures 1.7 and 1.8. Furthermore, within galaxies there are stars of different spectral types, some of them hot enough to emit lines similar to those observed in quasars, but there are also cool stars and interstellar medium with clouds of gas and dust that also emit narrow lines. This variety of emission sources complicates the task of classifying and differentiating quasars from galaxies, and even some stars, making it a complex problem to address in astronomy.

1.6 Motivation behind our study

As shown in the preceding sections, there remains a continued interest in developing classifiers capable of more accurately distinguishing between each spectral type present in the data released by machines like DESI. Rapid classifiers, such as RedRock, are limited to roughly matching spectra with a sample of templates that fail to capture the high variability or entirety of the physical phenomena present in stars, quasars, and galaxies. Consequently, achieving high precision in classifying each of these objects is challenging.

On the other hand, QuasarNET operates by identifying specific spectral lines, using their presence and intensity to determine whether an object is a quasar or not. However, the presence and intensity of these emission lines vary depending on whether we are observing the torus around the central region of a quasar, where we find a wide range of excited elements due to interactions with other particles at high speeds and interactions with X-ray emission from the central region. Conversely, if we observe a quasar right in its central zone, we will find broad lines in the spectrum due to multiple redshifts of the elements near the accretion disk, in addition to emission lines from highly ionized elements also found in the torus.

Furthermore, galaxies can also exhibit many of the aforementioned emissions in their

spectra, see Figure 1.7 and Figure 1.8, originating from the stars that comprise them, or they may even be confused with the spectrum of a star hot enough to exhibit emission lines of these highly ionized elements. This poses a challenge to achieving truly reliable classification based solely on emission lines within the spectra. Hence, we propose to employ more general and robust features, such as the patterns and spectrophotometric values that will be sought in this article.

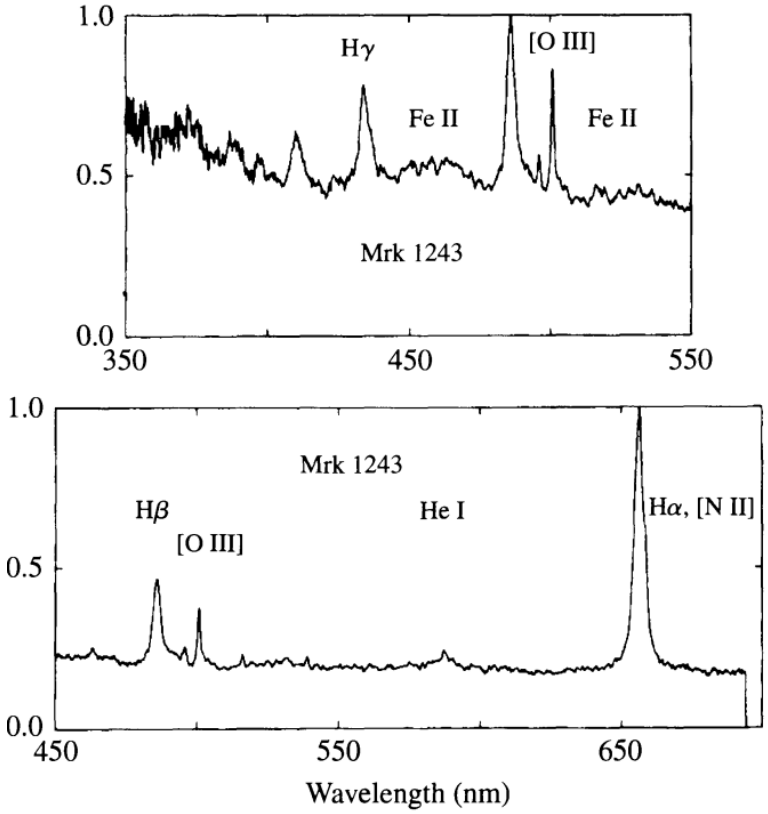

Figure 1.7: Visible Spectrum of Mrk 1243, a Seyfert I galaxy. This spectrum closely resembles the spectra of QSO A and QSO C shown in Figure 1.6, suggesting potential direct observation from the active nuclei. Notably, it exhibits broad lines characteristic of elements such as Fe[II], Hγ, Hα, and other ionized elements, indicative of high speeds within the elements. Extracted from (Ostlie & Carroll 1996).

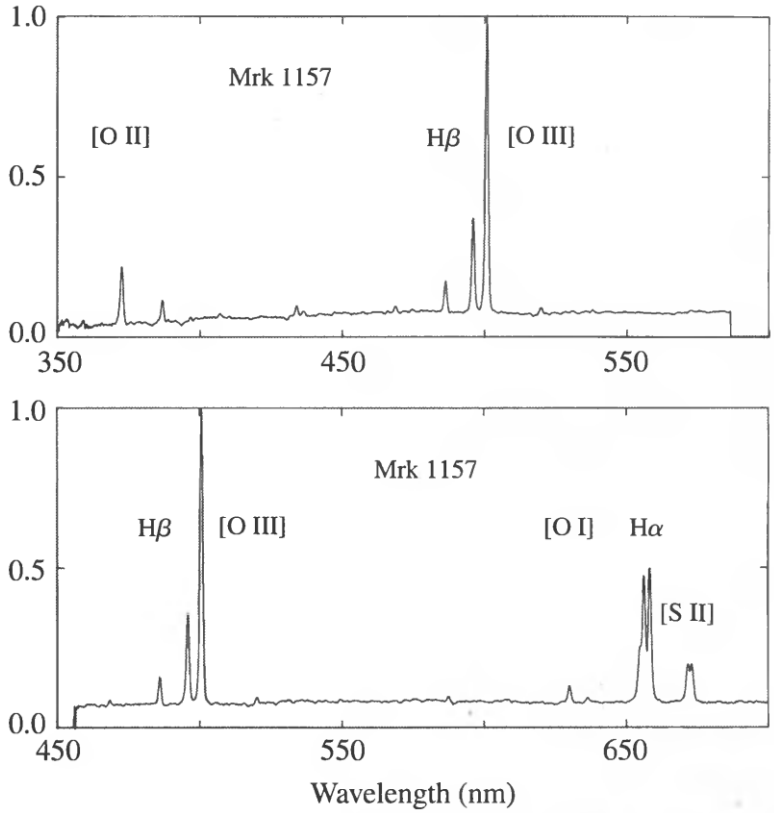

Figure 1.8: Visible Spectrum of Mrk 1157, a Seyfert II galaxy. This spectrum bears a striking resemblance to the spectrum of QSO B depicted in Figure 1.6, hinting at potential direct observation from the torus of the quasar. Noteworthy are the narrow lines evident, characteristic of elements like O[II], O[III], N[II], and S[II], suggesting emission originating from the dust within the torus. Extracted from (Ostlie & Carroll 1996).

In this way, our research focuses on addressing the scientific challenge of dealing with ambiguity in the classification of galaxies and quasars, while identifying new features to facilitate the creation of simple models for the classification of galaxies, quasars, and stars using spectrophotometric data from DESI. Our specific objectives are outlined as follows:

First, we aim to identify the differences and patterns within our spectrophotometric features that provide the most accurate predictions of spectral type for galaxies, quasars, and stars from the DESI's data. Our approach not only emphasizes precision but also aims to minimize the complexity of the neural network, ensuring efficiency without compromising effectiveness.

Second, leveraging the capabilities of the H2O tool, we will conduct rigorous testing to

identify optimal machine learning models for the classification task. Through systematic experimentation, we seek to refine our selection process, determining the models that exhibit the highest performance metrics and robustness while maintaining minimal complexity, that is, prioritizing the simplest model.

Finally, we will undertake a comprehensive comparative analysis. This involves juxtaposing the predictions generated by our newly developed spectrophotometric classifiers with those generated by existing DESI classifiers. By evaluating the concordance and disparities between the two sets of predictions, we aim to gain insights into the effectiveness and advancements offered by our approach compared to the established methodologies employed by DESI.

Chapter 2

Data

2.1 Specifics of the DESI data set

The treatment and description of the Early Data Release were carried out systematically. Initially, five primary target classes were identified: Milky Way objects (MWS), bright galaxies (BGS), luminous red galaxies (LRG), emission line galaxies (ELG), and quasars (QSO). Additionally, an *others* class was designated for any remaining targets.

Moreover, these classes were grouped according to stages of data combination, such as *per exposure*, representing data from a single exposure on the telescope; *Per night*, involving the combination of spectra obtained from multiple exposures during the same night; *Per tile*, encompassing the merging of spectra from the same mosaic, even if acquired on different nights; Finally, *per healpix*, denoting the combination of spectra based on their sky position, potentially spanning multiple nights and mosaics.

The data within the DESI project were systematically categorized into various catalogs [1]. The CMX catalog, specifically employed during the commissioning phase, played a pivotal role in collecting data in this preliminary stage. It acted as a repository for observations conducted during the instrument's commissioning, encompassing test observations, reference

[1] All of DESI's catalogs are published and publicly accessible in this virtual repository: https://data.desi.lbl.gov/public/edr/spectro/redux/fuji/healpix/sv3/

measurements, and other relevant data. The CMX catalog aimed to facilitate a comprehensive understanding of the instrument's performance, enabling scientists and engineers to address and fine-tune any technical issues before transitioning into subsequent phases of the DESI project.

During the target selection validation phase (SV1), central aspects included the determination of effective exposure times (*EFFTIME_ETC*) and the establishment of minimum observation specifications with (*EFFTIME_SPE*). SV1 marked the initial steps of Survey Validation (SV), providing essential data to validate and calibrate the instrument under actual observational conditions. This set the stage for the comprehensive exploration of the universe in subsequent DESI project phases.

The subsequent catalogs played distinctive roles in advancing the DESI project. SV2, focused on operations testing, played a role akin to a "dress rehearsal" for SV3. SV3, the One-Percent Catalog, marked a significant milestone, validating final operational procedures and compiling extensive samples for clustering studies. These samples, representing only 1% of the total data, boasted high fiber assignment completeness.

Concluding this sequence, the *special* catalog was dedicated to observations related to secondary or test targets, providing specialized insights. On the other hand, the *main* catalog formed the backbone of the DESI project, encapsulating data from the primary DESI survey starting with Data Release 1 (DR1). It comprised the bulk of the survey's observational data, offering a comprehensive and foundational dataset for addressing the overarching scientific goals of DESI. In Table 2.1, the number of primary objects per catalog is shown.

SURVEY	N_{BGS}	N_{ELG}	N_{LRG}	N_{QSO}	N_{STAR}	N_{SCND}
cmx	247	761	1,037	275	468	0
sv1	134,419	111,692	66,161	29,839	163,254	60,430
sv2	46,628	12,308	22,151	11,032	10,506	0
sv3	253,915	312,790	137,317	34,173	295,232	75,947
special	925	3,866	3,588	3,045	867	3,482
Total	428,758	437,664	227,318	76,079	466,447	137,148

Table 2.1: Count of primary objects in the DESI catalogs.

Targets were also grouped by observing conditions: *dark* for Dark nights, optimizing observations for luminous red galaxies (LRG), emission line galaxies (ELG), and QSOs. *bright* for Bright time, allocated for Bright Galaxy Survey (BGS) and Milky Way Survey (MWS) objects. The bright vs. dark distinction is based upon survey speed, or how quickly the instrument can accumulate signal-to-noise given the current observing conditions. Finally, *backup* for the worst observing conditions, typically resulting in noise, and *other* for none of the above categories.

This strategic organization facilitated a comprehensive and structured approach to managing and categorizing data during the Early Data Release, streamlining subsequent analysis and interpretation processes.

2.2 Choosing reference targets for the training and validation steps

In our data selection process, we opted for spectra corresponding to the sv3 catalog, as they represent the most reliable measurements with high completeness. Specifically, we focused on both dark and bright observing conditions, as they are optimal for studying various subtypes of galaxies and quasars. However, we excluded the backup category due to the lower quality of the measurements associated with it.

Moving on to the objectype label in the DESI catalogs, this attribute categorizes objects into three possible values. Firstly, *sky* refers to reference objects that serve as benchmarks. Secondly, *target* is assigned to objects linked to a fiber in the spectrograph. It's essential to note that discarding spectra labeled as *bad* is crucial in enhancing the training process, as these entries exhibit poor or low-definition measurements.

The negative values in the targetid field hold significance in our data analysis. Negative targetids indicate objects for which no specific target was assigned, often corresponding to ancillary measurements or those without a predefined target. For a more focused and accurate

analysis, we opted to discard entries with negative targetids from our dataset. Figure 2.1 displays spectra from each spectral type.

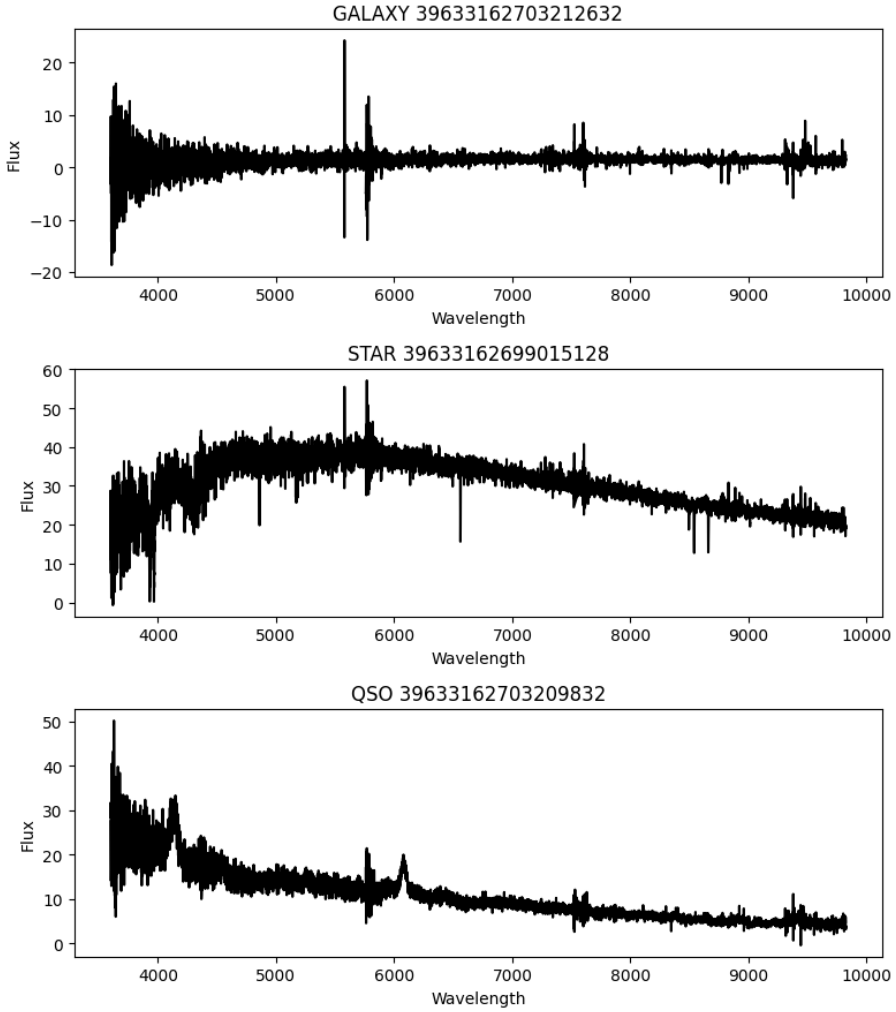

Figure 2.1: Spectra of a galaxy, star, and quasar measured by the DESI project. This is flux plotted against wavelength.

In total, we have 1,274,157 spectra suitable for training our models. Among these, 574,830 objects are from the Bright catalog, and 699,327 are from the Dark catalog. Within this dataset, there are a total of 964,814 galaxies, 51,221 quasars, and 258,122 stars.

The training process was conducted using 60% of the total data for each class, ensuring a stratified approach. Validation was performed using the remaining 40% of the data. Each object counts with the following information: normalized flux and wavelength values, a unique

code identifying each element, redshift along with its corresponding error, and astronomical coordinates in right ascension and declination.

Our training process employed a meticulous cross-validation strategy to assess the generalization capacity of our machine learning models. This technique partitions the dataset into multiple subsets, training the model on one part and evaluating it on another.

Chapter 3

Tools

3.1 H2O-3

H2O-3 is a powerful and versatile open-source software platform designed for machine learning and deep learning tasks. It offers a wide range of functionalities that make it a valuable tool for data scientists and researchers alike.

Primarily, H2O-3 serves as a platform for building and deploying machine learning and deep learning models. It supports both supervised and unsupervised learning techniques, allowing users to tackle a variety of tasks, from classification and regression to clustering and anomaly detection. One of its notable features is its ability to handle large datasets efficiently, making it suitable for big data applications.

In terms of model building, H2O-3 provides a rich set of algorithms, including popular ones like Random Forest, Generalized Linear Models (GLM), Gradient Boosting Machines (GBM), and Deep Learning. This diversity enables users to experiment with different algorithms and select the ones that best suit their data and objectives. Additionally, H2O-3 offers automated model selection capabilities, allowing users to test multiple models and recommend the most suitable one based on performance metrics.

Moreover, H2O-3 provides seamless integration with Python, allowing users to leverage

its capabilities within the Python ecosystem. This integration facilitates easy adoption for Python users and enables them to incorporate H2O-3 into their existing workflows effortlessly.

It's important to note that H2O-3 is primarily designed for tabular data, meaning it works well with structured datasets represented as tables. While this limitation precludes the direct use of spectral flux arrays as inputs, we can address this by transforming the extracted features from Chapter 4 into tabular format. This ensures compatibility with H2O-3 and enables us to leverage its powerful modeling capabilities.

To utilize H2O-3 effectively, input data needs to be converted into the H2O frame format. An H2O frame is a distributed and parallel in-memory representation of a dataset that allows for efficient processing and analysis. To convert data into an H2O frame, one typically uses the H2O Python API or other H2O-compatible interfaces. This process involves loading the data into memory and converting it into the H2O frame format using appropriate functions as $h2o.H2OFrame()$. By following this procedure, we can seamlessly integrate our data into H2O-3 and leverage it for our classification tasks.

In this manner, H2O-3 will be employed to comprehensively evaluate the features extracted from the spectra across various models, selectively choosing those showcasing optimal performance. Following this, we will engage in the meticulous construction and fine-tuning of the chosen models to address the demanding tasks of spectral classification and redshift estimation. This thorough approach will guarantee that our models are thoroughly optimized and well-prepared to handle the complexities of astronomical analysis, delivering precise and insightful outcomes in our research pursuits.

3.2 SAM (spectra analyzer machine)

The SAM (Spectra Analyzer Machine) neural network has been meticulously crafted with dual objectives: the classification and redshift measurement of celestial objects. It serves as a paradigmatic example of the application of advanced machine learning techniques, notably

Convolutional Neural Networks (CNNs), in the field of computational astronomy (Cárdenas & Fajardo 2023).

CNNs are specialized neural networks tailored for extracting abstract patterns from data. SAM's architecture is firmly rooted in this technology, featuring four convolutional layers with filters meticulously crafted to extract pertinent features from the spectra. Each convolutional layer, comprising 64, 128, 256, and 256 filters respectively, conducts convolution operations and applies activation functions such as the Rectified Linear Unit (ReLU) to introduce non-linearities into the network (Cárdenas & Fajardo 2023). Additionally, SAM incorporates max-pooling reduction in each layer to reduce model complexity, enhancing and expediting the training process. Furthermore, dropout is employed to turn off the 50% of the neurons in the subsequent layer, look at Figure 3.1, preventing neuron weight overfitting and bolstering the network's predictive capability.

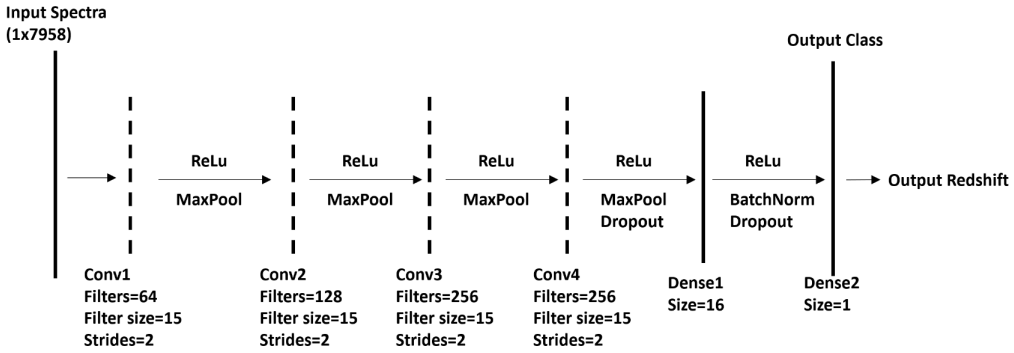

Figure 3.1: Architecture of SAM for the classification exercise. Extracted from Cárdenas & Fajardo (2023)

SAM demonstrates remarkable efficiency in identifying and extracting patterns from incoming spectral features. This proficiency empowers the network to adeptly differentiate between categories such as galaxies, quasars (QSO), and stars, rendering it an invaluable tool for automating classification tasks within the extensive astronomical dataset collated from the *dark* and *bright* programs of DESI.

Following rigorous training and evaluation, SAM achieved an impressive overall accuracy of 98.41% as evidenced by the confusion matrix, underscoring its capacity to accurately

classify the majority of astronomical spectra, as seen in Figure 3.2. Moreover, it attained an F1 Score of 0.9824, indicative of a balanced performance between precision and recall, thereby mitigating both false positives and false negatives (Cárdenas & Fajardo 2023).

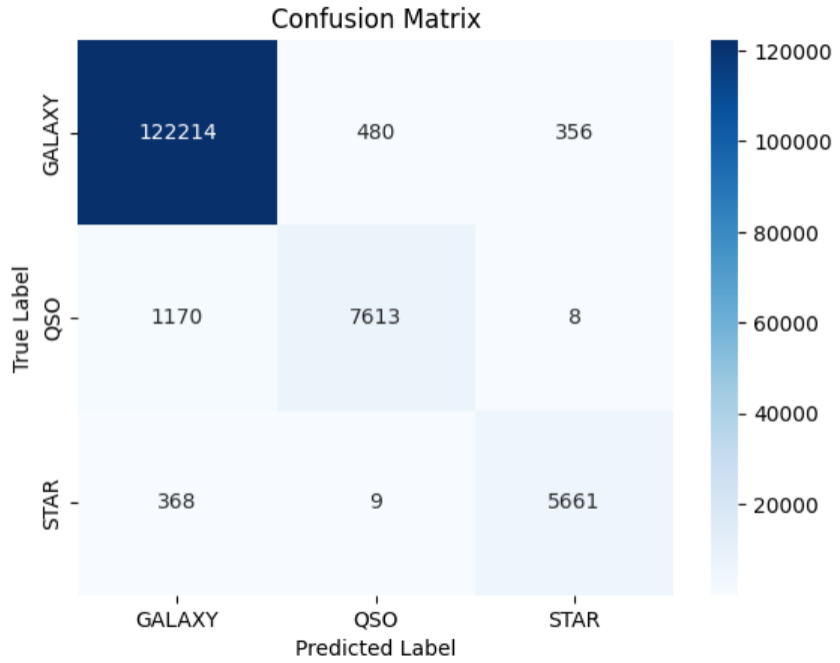

Figure 3.2: SAM's confusion matrix for spectral classification. The matrix displays real spectral types in the rows and predicted classes in the columns. Correctly classified objects are represented by elements along the diagonal. The blue color scale indicates the density of elements classified within each cell. Extracted from (Cárdenas & Fajardo 2023).

However, a crucial aspect of our work entails fine-tuning SAM once again, ensuring that it learns to interpret the forthcoming features in the most efficient and beneficial manner for predicting the spectral type and redshift of galaxies and quasars, as detailed in the fourth section. This entails optimizing the hyperparameters of the network, including the number of convolutional layers, the number of neurons, filter sizes, activation functions, and other aforementioned functions.

Chapter 4

Methodology

4.1 Data compilation

As mentioned in Chapter 2, we utilize the bright and dark catalogues of the DESI data. Subsequently, we extracted all data from the public virtual repository[1]. Now, we will elucidate the structure of this virtual repository.

Firstly, once within the repository, our focus lies on the directory *public/edr/spectro/* as it houses all spectra released by the Early Data Release (EDR). Here, we encounter templates and calibration spectra. However, our pursuit is for the redux spectra. Hence, we navigate to *public/edr/spectro/redux/fuji/*. Depending on our interests, various catalogues are available, categorized as per exposures, tiles, and healpix.

Additionally, notable zcatalogues, such as the *zall-pix-fuji* catalogue, contain comprehensive compilations from all catalogues, programs, and classes detailed in Chapter 2. For our purposes, we utilize the healpix catalogue, within which are the *cmx*, *special*, *sv1*, *sv2*, and *sv3* catalogues.

Consequently, our target being the *sv3* catalogue, the final path is *public/edr/spectro/redux/fuji/healpix/sv3/*. Here, we access the *bright* and *dark* directories.

[1] DESI's virtual repository: https://data.desi.lbl.gov/public/edr/spectro/redux/fuji/healpix/sv3/

Now that the catalogues are located, comprehending how to extract data from them is imperative. We will discuss the *dark* directory, though the procedure is identical for the *bright*.

We observe a hierarchical structure of directories labeled with numbers such as *100/*, *101/*, *102/*, and so forth, constituting the first layer. It is essential to traverse through all of them. Within each first-layer directory, we encounter a subsequent set of directories labeled with five digits, e.g., *28031/*. This constitutes the second layer, requiring traversal of all directories within. Eventually, within one of the second-layer directories, a collection of fits, csv, and h5 files constitutes the third layer of the structure. From the third layer, we search for the coadd and redrock fits files, containing fluxes in the B, R, Z DESI's filters and specific information of each spectrum such as spectral type, respectively. Notably, each pair of fits files (coadd-redrock) contains identical spectra; if the coadd file contains the fluxes of spectra 1, 2, 3, then the redrock files contain data of spectra 1, 2, 3.

```
Filename: DataDESI_691_752.fits
No.    Name      Ver    Type      Cards   Dimensions   Format
  0  PRIMARY       1 PrimaryHDU       4   ()
  1                1 BinTableHDU     26   89878R x 9C   [7A, D, D, 7A, D, 21A, D, D, 7A]
  2  B_FLUX        1 ImageHDU         8   (2751, 89878)   float32
  3  R_FLUX        1 ImageHDU         8   (2326, 89878)   float32
  4  Z_FLUX        1 ImageHDU         8   (2881, 89878)   float32
```

Figure 4.1: The structure of one of the ten large fits files housing all of DESI's data. Here, both the BinTable HDU and Image HDU elements can be observed, containing tabular labels and fluxes, respectively.

In total, 373 coadd-redrock pairs are found within the bright directory, and 379 pairs within the dark directory. To facilitate data manipulation, all spectra are compiled into 10 fits files. Each of these large fits files occupies approximately 4.1 GB of memory, containing a total of 1,274,157 spectra.

Regarding the structure of these 10 large fits files, we maintain the original structure provided by DESI. This includes one BinTable HDU element in the first position, containing tabular data of the spectra. Additionally, three image HDU elements store flux data for the three observation ranges - B, R, and Z, as shown in Figure 4.1. We have preserved labels such

as spectral type, object type, morphology type, redshift, its calculated error, right ascension, declination, and others. This is illustrated in Figure 4.2 below:

SPECTYPE	Z	TARGETID	OBJTYPE	Z_ERR	SUBTYPE	TARGET_RA	TARGET_DEC	MORPHTYPE
str7	float64	float64	str7	float64	str21	float64	float64	str7
GALAXY	0.26495553753651446	3.9632986768936136e+16	TGT	7.675386768239847e-05	0.0	251.13412200880515	35.23854486920474	SER
STAR	-0.00017170741474576307	3.963298676474267e+16	TGT	6.297295578494421e-06	G	250.85477816221342	35.164102405409004	PSF
GALAXY	0.1138600133189154	3.963298676893551e+16	TGT	7.161186384705142e-05	0.0	251.09286363782294	35.15489017306431	SER
STAR	-0.0005461945376769534	3.9632981823850696e+16	TGT	5.1880335838730204e-06	F	251.12491082094607	35.1200088112371	PSF
STAR	-0.0019956912923479522	6.160937341272732e+17	SKY	4.1311493573349107e-48	CV	251.0223884874211	35.123503519956884	0.0
GALAXY	0.16752992864325952	3.963298676893413e+16	TGT	7.170038677014659e-06	0.0	250.99738588331317	35.211964245082584	SER

Figure 4.2: The table depicts one of the ten large fits files containing DESI's data. Each column represents a distinct label, while each row corresponds to a different spectrum. Each row aligns with specific flux values located at the same index in the Image HDU elements.

4.2 Data standardization

After downloading all the data, we proceeded with data standardization. This process is essential and highly beneficial in machine learning workflows. By standardizing the data, we bring it to a common scale, making it easier for models to discern meaningful relationships among the features. This uniformity eliminates biases that may arise due to differing scales of the features. Consequently, models can enhance their prediction capabilities.

To standardize the data, we calculate the average values and standard deviations for each flux value at every wavelength across all spectra. Essentially, this involves standardizing each flux value independently while considering the distribution of all other spectra values at the same wavelength. In other words, for each flux value at a specific wavelength, we subtract the mean of all flux values and then divide by the standard deviation of all flux values at that particular wavelength.

The standardization formula is represented as follows:

$$z_i = \frac{x_i - \bar{x_i}}{\sigma_{x_i}}.$$

Where z_i represents the standardized value of flux, x_i stands for the original flux value,

$\bar{x}_i$ denotes the mean flux value across all spectra, and σ_{x_i} indicates the standard deviation of flux across all spectra. All these calculations are carried out for a specific wavelength denoted by the subscript i.

4.3 Deprecating incomplete and bad spectra

Despite having already discarded spectra labeled as *bad* by this point, and having taken the best measurements reported by DESI, we have found within the spectra some that exhibit anomalous behaviors, data absence, or excessive noise within one of DESI's three B, R, Z filters. We consider that incorporating these within the models' training could introduce biases or confuse the pattern recognition associated with each spectral class.

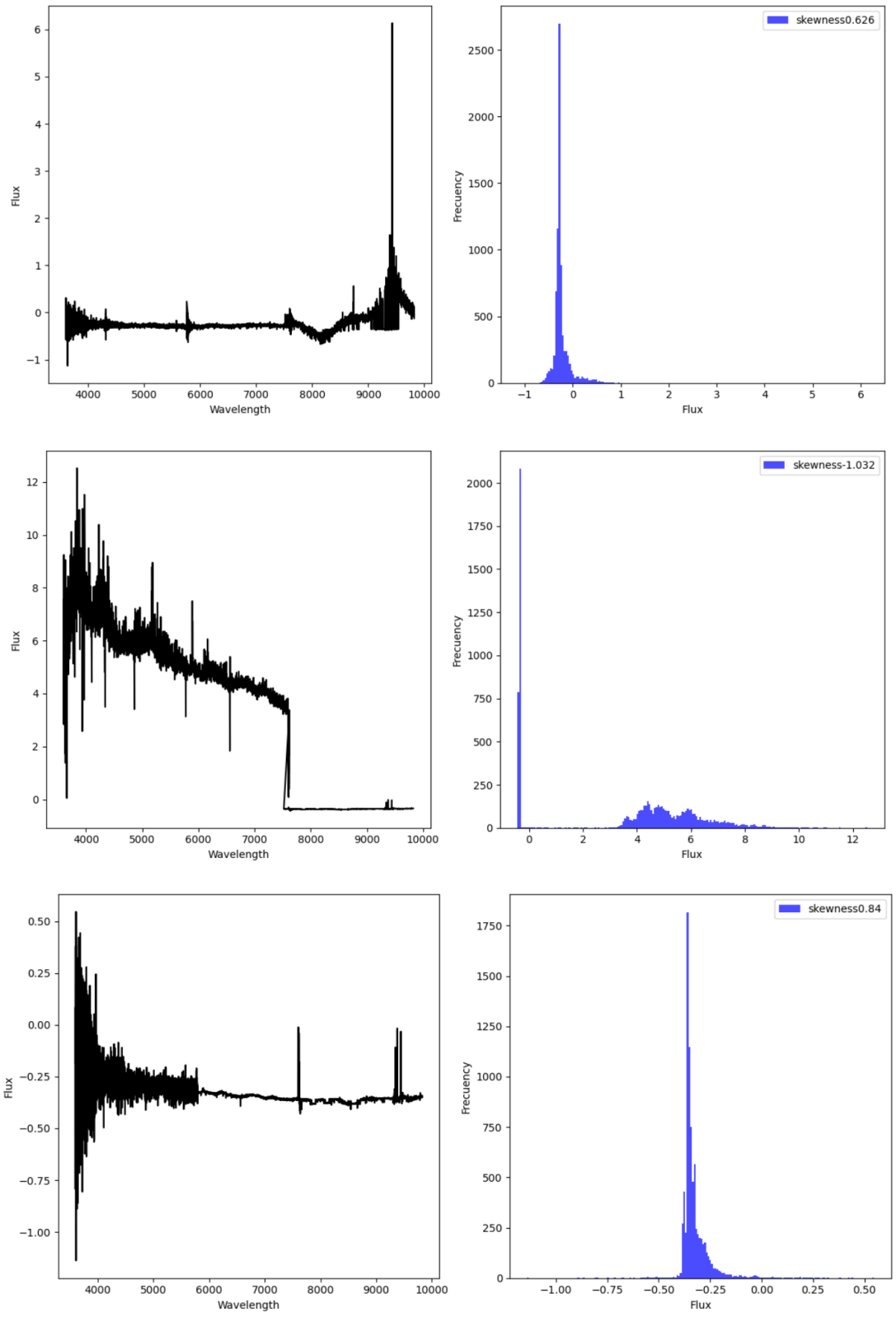

Figure 4.3: Spectra with measurement errors within the DESI catalog due to noise saturation, data absence, or anomalous behaviors. In the left column, the images of these spectra appear, and on the right, their respective flux histograms. In these histograms, flux saturation within the range [-0.5, -0.25] is observed.

The method we used to identify around 2,538 objects with this behavior involved dividing each spectrum's flux data into 6 ranges, where we calculated the MAD parameter for each of these. This is the same parameter described in the Model inputs section, as it provides information on how much the flux data varies in those regions. Given that bad spectra exhibit noise or data absence in certain regions, there is not much flux variation here, resulting in a small MAD. After this, spectra were labeled if any of the 6 MAD values were less than or equal to 1% of their maximum MAD value. However, we noticed that quite a few good quasars were slipping through among the bad spectra, as this behavior is not uncommon for this class.

Therefore, we decided to use an additional requirement to properly filter out the spectra we were interested in discarding. From the flux histograms corresponding to each spectrum, we noticed that bad spectra saturated the flux within the range of values [-0.5, -0.25], and thus their histograms exhibited high peaks there, with values greater than 1,400 counts.

Thus, spectra that met these two requirements appeared as those in the Figure 4.3, and therefore were discarded for the training process.

4.4 Model inputs: Features and data

A significant part of this work revolves around determining the inputs that will best serve the models, aiming to strike a balance between reduced complexity and maximum classification metrics. To achieve this, a set of spectrophotometric features, different from those reported in the literature thus far, has been devised. These features estimate various aspects of quasars, galaxies, and stars, including their temperature, brightness magnitude in different bands, colors, and the behavior of their spectral data.

Additionally, for the training and fine-tuning of the SAM tool, utilizing the complete flux array of each spectrum as input has been proposed. Given the notable importance of flux in the u-band within the spectrophotometric features, it will be compared with another

model created by SAM that solely receives flux data within the u-band. Below, each of the aforementioned features is detailed.

4.4.1 Spectrophotometric features

Photometric band fluxes

A photometric band, in astronomy, refers to a specific range of wavelengths within the electromagnetic spectrum used to measure the brightness of celestial objects. These bands are crucial for analyzing the characteristics and properties of stars, galaxies, and quasars.

In our study, we opted not to utilize the color bands provided in the DESI flux data, namely B (3600-5800) Å, R (5760-7620) Å, and Z (7520-9824) Å, due to their considerable width encompassing various regions of the ultraviolet, visible, near-infrared, and far-infrared spectra, leading to significant information loss. Instead, we chose to employ the filters from The Sloan Digital Sky Survey Photometric System [2]. This system defines the bands u, g, r, i, and z with wavelength ranges of (3055.11, 4030.64) Å, (3797.64, 5553.04) Å, (3418.23, 6994.42) Å, (6692.41, 8400.32) Å, and (8385, 9875) Å, respectively.

The Sloan Digital Sky Survey (SDSS) Photometric System provides a well-established framework for photometric measurements. We selected these bands because they offer a more precise representation of how the flux vary across the spectrum of galaxies, quasars, and stars.

By partitioning the spectrum into these defined band regions, denoted as *ugriz* since now, we calculated the total flux within each wavelength range for different spectral types, galaxies, quasars, and stars. Consequently, we obtained a set of five features, one for each band in the *ugriz* system, representing energy magnitudes. These features are labeled in our codes[3] as *FLUX_U*, *FLUX_G*, *FLUX_R*, *FLUX_I*, *FLUX_Z*. The Figure below shows the distribution of the *FLUX_U* values, the distributions of the others can be found in Appendix A, as they are

[2] The Sloan Digital Sky Survey Photometric System: http://svo2.cab.inta-csic.es/svo/theory/fps3/index.php?id=SLOAN/SDSS.g

[3] Virtual GitHub repository: https://github.com/DanielFajardo1/Mapping-the-Universe.git

all very similar:

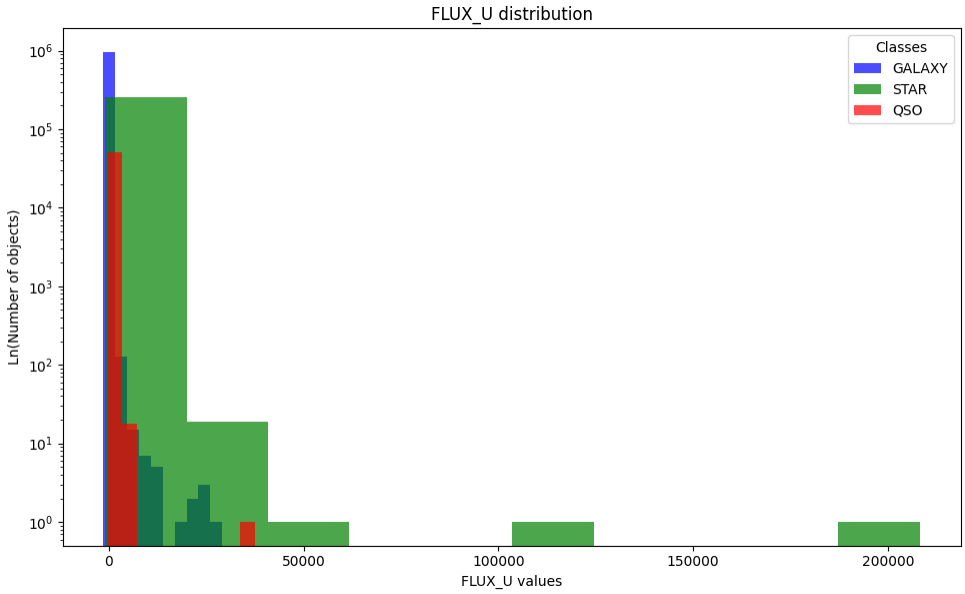

Figure 4.4: *FLUX_U* values distribution for each one of the three spectral types. Here it's important to note that this distribution suggests making a cut to only consider the outer parts of it, the wings of the distribution, as a key parameter for the classification of stars and quasars that deviate from the values where most galaxies lie.

Synthetic band fluxes

Similar to the fluxes across the photometric bands ugriz mentioned earlier, four synthetic filters were crafted. Two of these are broad, assessing the total flux in small and large wavelength regions, while the other two are narrow bands centered on zones where prominent emission lines occur, primarily in galaxies and quasars, but with generally distinct magnitudes. These four synthetic filters are denoted as F0, F1, F2, and F3, with wavelength ranges of (3055.11, 3756) Å, (5578, 5582) Å, (7601, 7603) Å, and (9310, 9570) Å, respectively. As described, these aim to discern patterns in the onset, termination, and energy content of the emission lines within the narrow bands.

To determine the values of these synthetic bands, we drew inspiration from the spectra resulting from the flux contributions of all galaxies, quasars, and stars, as well as the average flux spectra of each of these three classes. This can be understood by referring to Figure 4.5,

where the differences between these spectra are clearly visible. Galaxies, quasars, and stars contribute differently to the total flux at each wavelength, and the mean values also vary significantly in certain parts of the spectra. These patterns guided us to delve deeply and precisely into these areas using the bands F0, F1, F2, and F3.

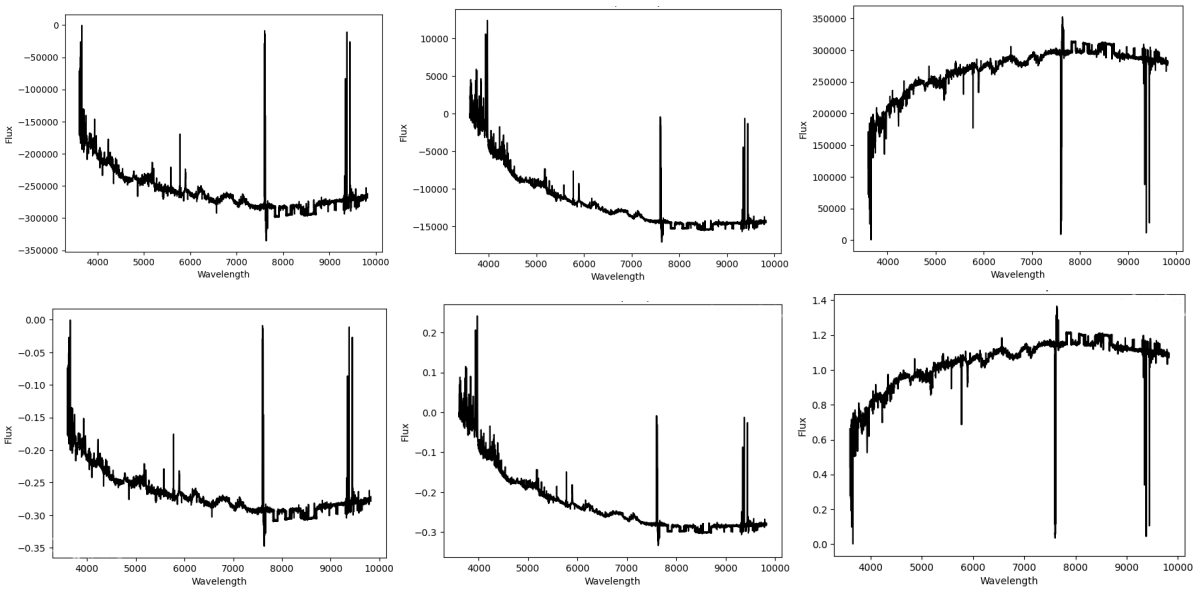

Figure 4.5: The three upper graphs depict spectra obtained from the flux contributions of all galaxies (upper left graph), quasars (upper center graph), and stars (upper right graph).The distinct shapes of these spectra at shorter and longer wavelengths are apparent. Additionally, the two prominent emission lines in the central region are noteworthy, with identical positions for quasars and galaxies but differing magnitudes. On the other hand, the lower set of graphs illustrates spectra constructed from the mean flux of galaxies (lower left graph), the mean flux of quasars (lower center graph), and the mean flux of stars (lower right graph). Again, differences in the values taken by the three spectra and the way in which they start and finish can be appreciated.

From these four synthetic filters, distributions were obtained where the tails of these distributions allow us to select galaxies and stars that deviate from the concentration point of quasars. These distributions are presented in Figure 4.6.

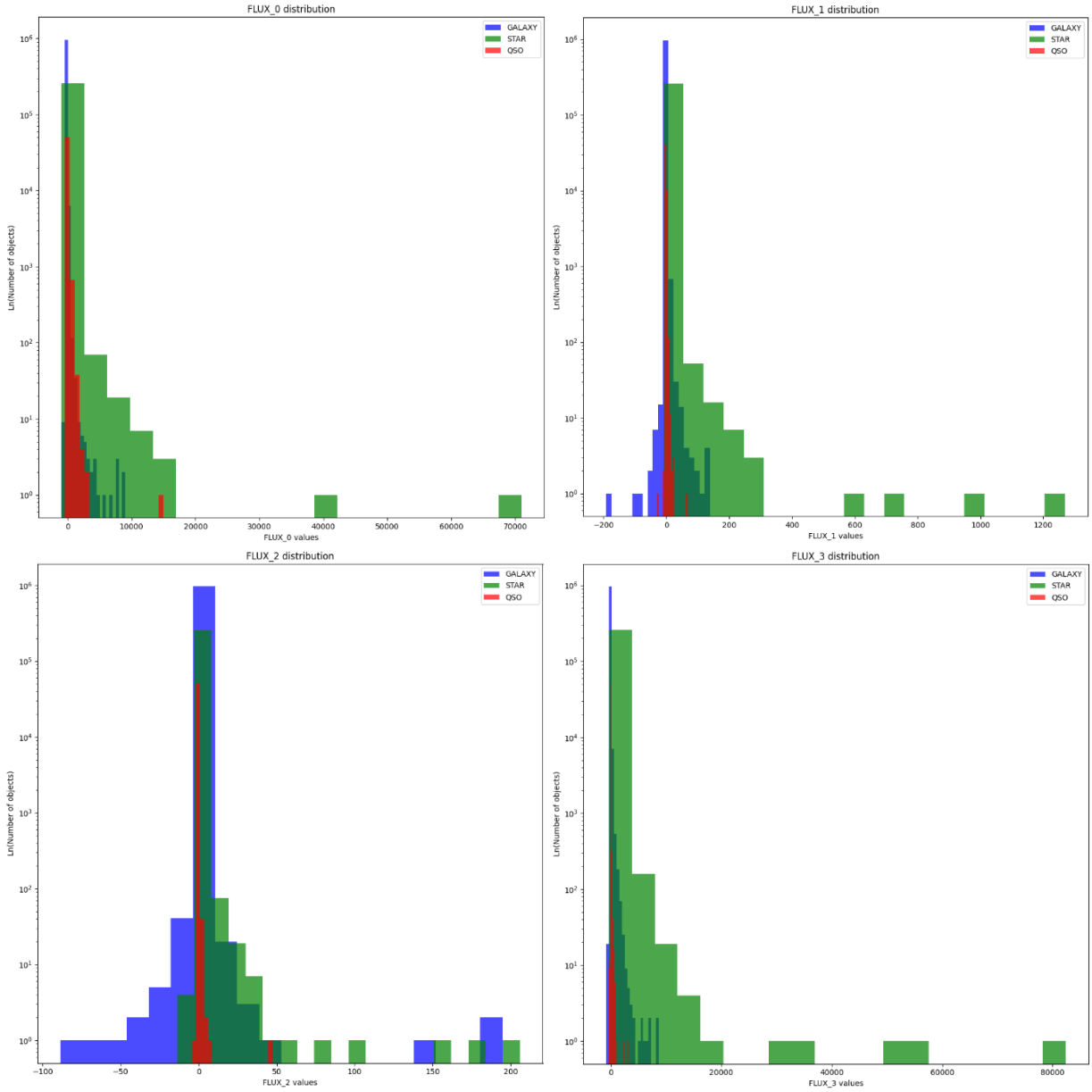

Figure 4.6: At the top of the graph, distributions of synthetic filters 0 are illustrated on the left, and 1 on the right. Meanwhile, at the bottom, distributions of filters 2 are shown on the left, and 3 on the right. The tails of these distributions provide an important parameter for distinguishing spectral classes.

Band fluxes slopes

From the total fluxes obtained from the ugriz bands and the four synthetic filters, certain trends were identified in the way targets were organized when considering them as coordinates.

These trends are effectively represented as slopes or rates of change of flux from one band to another. As depicted in Figure 4.7, classes are organized along lines with different slopes, notably distinguishing quasars from stars. Therefore, some of these slopes were taken as additional parameters. These slopes are labeled as *U/G*, *R/U*, *R/Z*, *G/Z*, *I/Z*, *U/Z*, *F0/G*, *F0/R*, *F0/I*, *F0/Z*, *F0/F1*, *F0/F2*, *F0/F3*, *F1/U*, *F1/R*, *F1/I*, *F1/Z*, *F1/F2*, *F2/R*, *F2/I*, *F2/Z*, *F2/F3*, *F3/U*, *F3/R*, *F3/I*, and *F3/Z*. Where *U*, *G*, *R*, *I*, *Z* are the same total fluxes *FLUX_U*, *FLUX_G*, *FLUX_R*, *FLUX_I*, *FLUX_Z*.

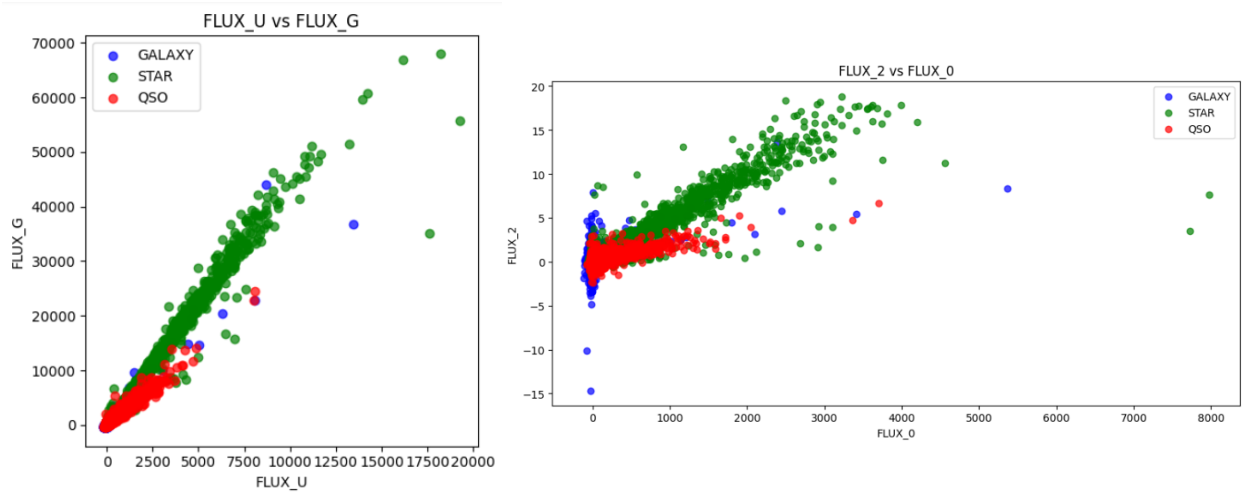

Figure 4.7: The graphs depict scatter plots of *FLUX_G* versus *FLUX_U*, and *F2* versus *F0*. It can be observed that galaxies, quasars, and stars tend to align along distinct lines with varying slopes. These two slopes correspond to *U/G* and *F0/F2*.

Color indexes

The color index in astronomy is a measure indicating the chromatic appearance of a celestial object, such as a star. It is used to describe how blue or red a star appears compared to a standard reference. In our study, we utilized variables of the total fluxes in the ugriz bands to compute color indices.

There are numerous possible combinations to derive color, involving the difference between two of all the calculated filters. However, many of these combinations are either redundant or fail to achieve a clear distinction among the spectral types we aim to classify. After discarding

such color indices that didn't provide useful information, we retained only the following: U-R, U-G, G-Z, R-Z, I-Z, and the slope (G-Z)/(R-Z).

In particular, we observed that the slope (G-Z)/(R-Z) exhibits a morphology enabling effective differentiation of stars from the other two spectral classes, thereby facilitating their identification through a circular mask.

Particularly, these features exhibit a significant dissociation in the distribution of elements among the three spectral types. In Figure 4.8, it is evident how galaxies and stars can be distinguished from quasars, whose values are highly concentrated around zero.

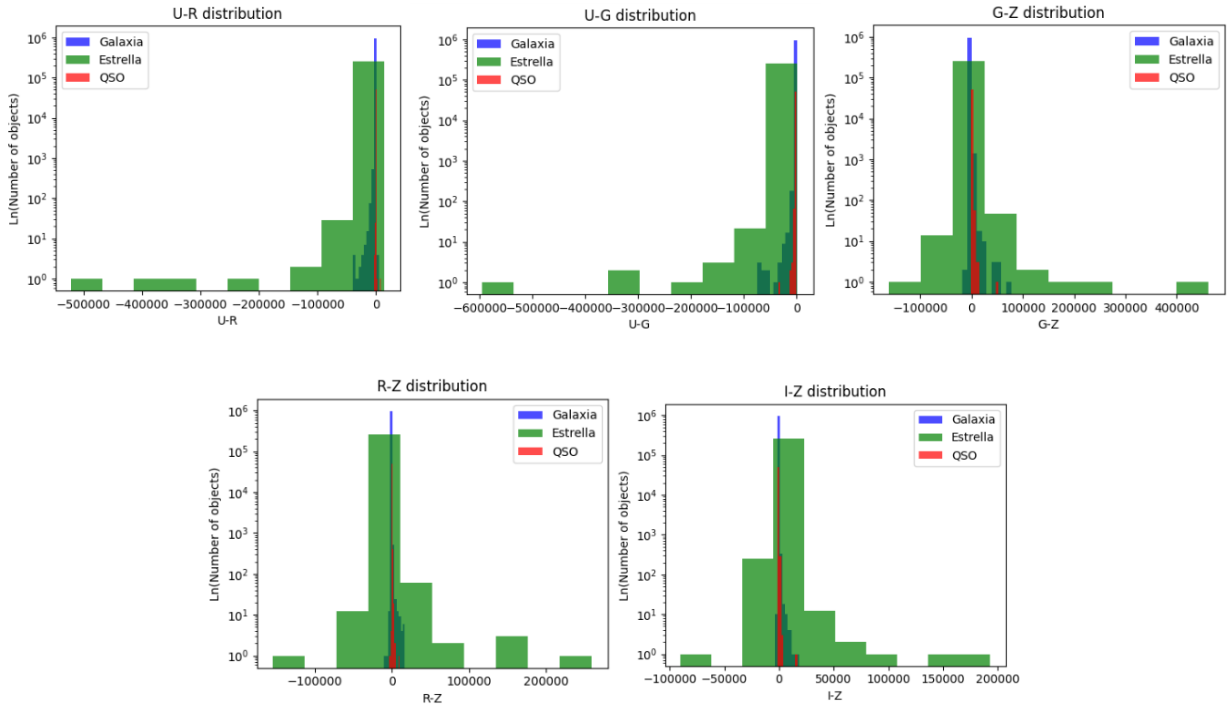

Figure 4.8: Distributions of each of the color indices taken as features for classification. It can be observed how quasars have a distribution highly concentrated around zero, while galaxies and stars spread over a wider range.

Magnitude modules and temperature

Once we have computed the total fluxes received for each of the bands, we proceed to extract the magnitude modules between each pair of fluxes, being careful not to be redundant. This is achieved through the relationship:

$$m_i - m_j = -2.5 \log \frac{FLUX\_i}{FLUX\_j} + const = a + \frac{b}{T_c}. \tag{4.1}$$

Here, i and j can take the values U, G, R, I, Z; m_i and m_j represent the brightness magnitudes of the target corresponding to the bands i and j used; $FLUX\_i$ and $FLUX\_j$ are the total fluxes within the spectrum covered by bands i and j respectively; T_c denotes the color temperature of the object, derived from the two fluxes used, while a and b are constants.

These quantities not only provide us with a measure of the temperature of each of the objects within the spectral types of galaxies, quasars, and stars, but it is also expected that the obtained values show a close relationship with current astronomical models. As mentioned in Chapter 1, quasars are active nuclei, and particles near both the accretion disk and those within the toroid are at high temperatures. Similarly, galaxies can also reach temperatures of several million degrees Celsius, making them significantly hotter than many stars.

This feature is particularly interesting because it could be used to carry out sub-classifications of quasars. It would be feasible to predict whether their toroids or nuclei are being directly observed before conducting a deeper analysis of the emission line morphology within the spectrum.

Then, insted of considering all the possible color temperatures, we decided to use the pairs $(i, j) = (U, G)$, (U, R), (G, Z), (R, Z), (I, Z). So, we count with five color temperature features. These five features are illustrated in Figure 4.9.

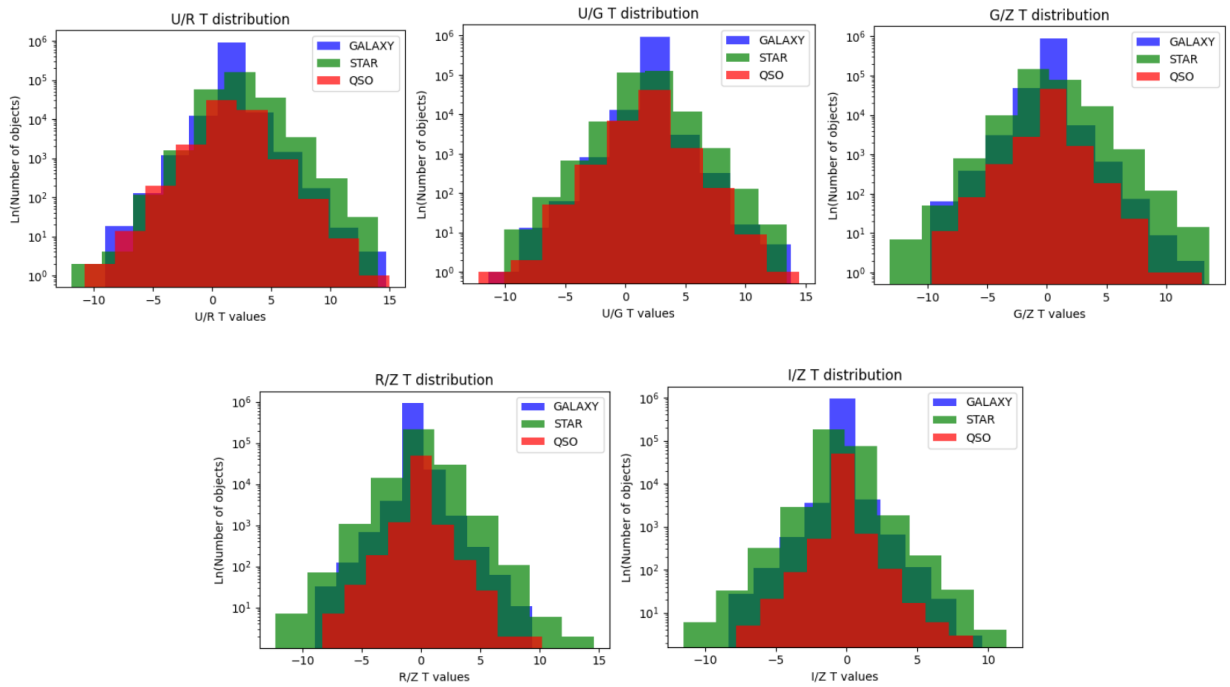

Figure 4.9: The distributions of each of the magnitude modules or color temperature, taken as features for classification, exhibit similar behavior among them. However, it is expected that the classification model will discover abstract patterns within them. The abbreviation U/R_T implies that this graph corresponds to the color temperature calculated from the logarithm of the ratio between the fluxes in the u and r bands. The same applies to the others.

Curvature

The spectra from Figure 4.5, and the curvatures of the spectra of galaxies, quasars, and stars suggested to us that there should be a certain trend in the value of these curvatures and whether they were positive or negative, we could provide an additional clue about what the spectral type of the observed spectrum was. For this, a simple fit was performed, of a polynomial of degree 2, whose form is ax^2+bx+c. Then, the curvature parameter is given by the second derivative of this fit, so curvature= 2a. This value was computed for all spectra and within our lines of code it corresponds to the label *CURV*, and their distributions are depicted in Figure 4.10.

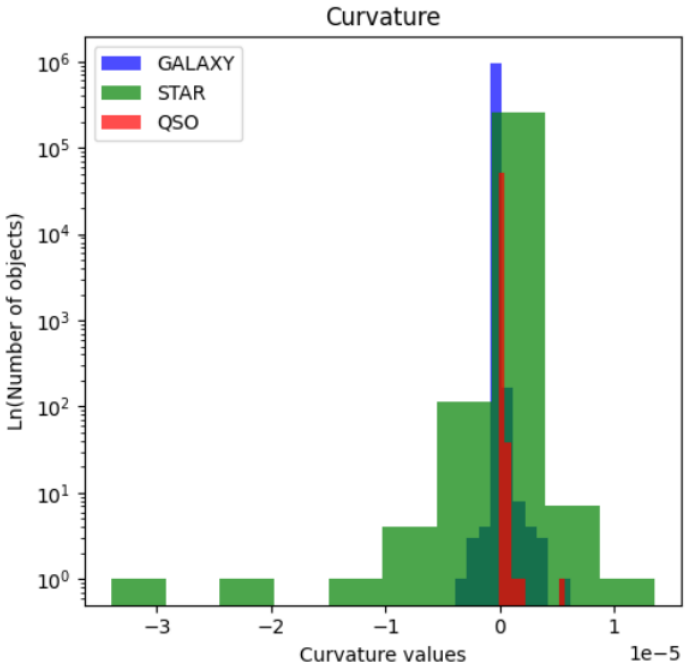

Figure 4.10: Similar to previous distributions, a clear distinction between the three spectral types is evident in the tails of the curvature value distributions. The x-axis is plotted on a logarithmic scale with base 5.

Median absolute deviation (MAD)

The MAD (Median Absolute Deviation) parameter is a robust measure of dispersion used to assess the variability within a dataset. Unlike the standard deviation, which is influenced by outliers, MAD employs the median as a central measure and evaluates the absolute distance of each data point from this median. MAD is calculated using the following formula:

$$MAD = median(|F_i - median(F)|)$$

Where F_i represents each of the spectral flux values, and $median(F)$ denotes the median of the flux values from the spectrum. This formula computes the median of the absolute differences between each data point and the median of the original dataset.

In summary, MAD serves as a useful metric for assessing the dispersion of spectrum flux values, particularly in the presence of outliers, as it is less sensitive to them compared to the

standard deviation. The corresponding distributions are illustrated below in Figure 4.11.

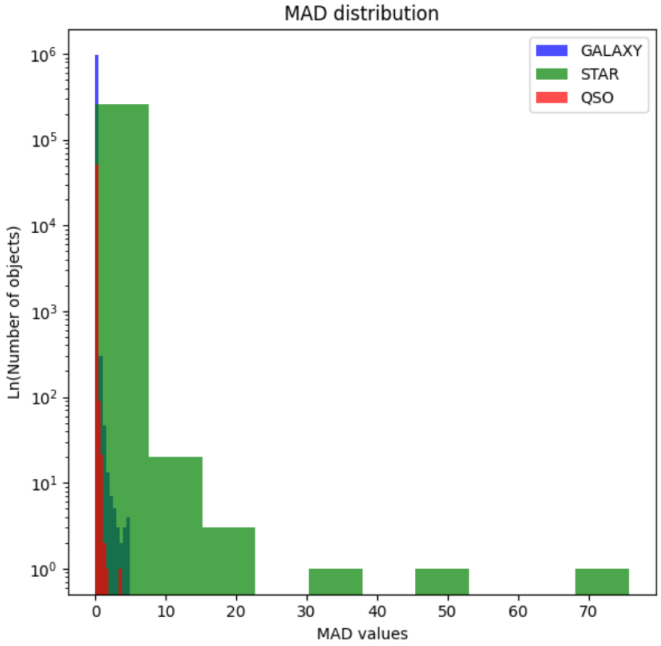

Figure 4.11: Distributions of the MAD parameter values calculated for the three spectral types are depicted. A higher dispersion in the flux values of stars is evident, while quasars and galaxies exhibit values that hover around zero.

Abbe value

The Abbe parameter serves as a metric to gauge the smoothness of a spectrum. It provides insight into how gradual or abrupt changes occur between adjacent wavelengths. To calculate this value, one compares the quadratic increment $(f_{i+1} - f_i)^2$ with the standard deviation of the spectrum. The formula for the calculation of this parameter is as follows:

$$Abbe = \frac{n}{2(n-1)} \frac{\sum_{i=1}^{n-1}(f_{i+1} - f_i)^2}{\sum_{i=1}^{n}(f_i - \bar{f})^2}$$

Where n is the total number of flux values contained in the spectrum, f_i is the i-th flux value of the spectrum, and $\bar{f}$ is the average flux value of the same. The distributions of these values is illustrated in Figure 4.12.

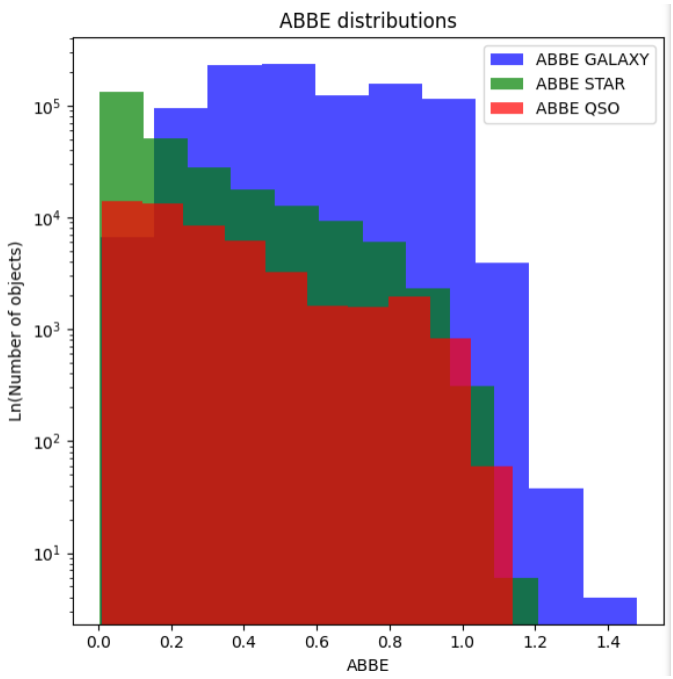

Figure 4.12: Distributions of the Abbe parameter values calculated for the three spectral types are depicted. Values close to 1 represent overly noisy spectra, whereas smaller values indicate smoother spectra.

χ^2 Templates

Inspired by the tactic used by DESI's fast classifier, RedRock, we used the three spectra from Figure 4.5, which correspond to the mean flux values of each class: galaxies, quasars, and stars, as templates. After this, for each spectrum, we calculated the chi-square homogeneity test to have a measure of how much this spectrum resembles the mean spectrum of each class. In this way, it is expected that a star spectrum resembles more the mean spectrum of stars, and similarly for quasars and galaxies. The way in which this chi-square parameter is calculated is:

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_i - T_i)^2}{T_i}$$

Where T_i is the i-th mean flux value of the template that we use, n is the number of flux values retained by the spectrum, and f_i is the i-th flux value of the spectrum. Because we

are using three templates, we have three different chi square features for each spectrum.

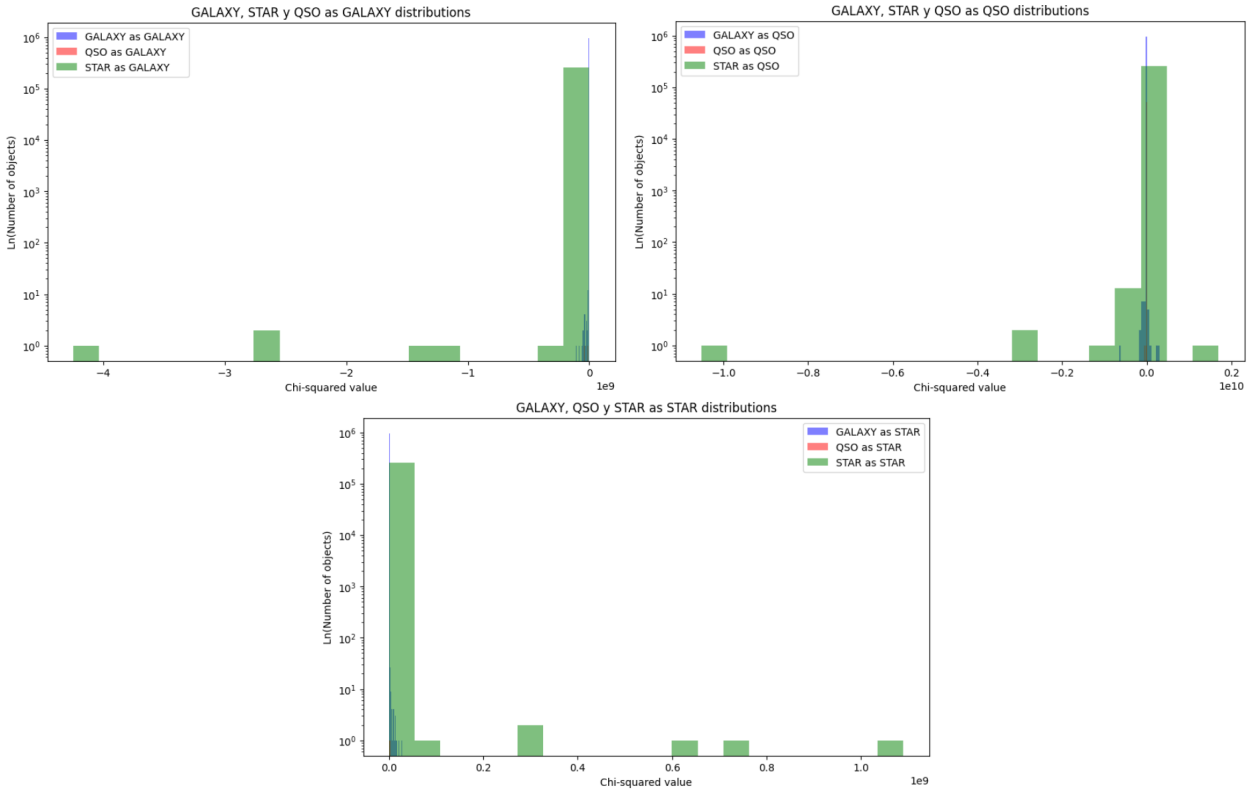

Figure 4.13: The distributions of each of the chi-square values from the three templates used. The top template contains the distribution of the chi-square values using the template of the mean flux of galaxies, while the middle one uses the template of quasars, and the bottom one uses the template of stars.In these graphs, the x-axis is in base 9 logarithm, except for the middle one, which is in base 10.

The distributions in Figure 4.13 exhibit quite a good behavior for distinguishing between spectral types once again, as very tight distributions are observed for quasars, allowing us to differentiate galaxies and stars from them.

In addition to these three features, we are also quite interested in the values obtained from the ratios between these three quantities, as distributions were found that made the differences between the three spectral types a bit more explicit. This other group of ratios are denoted as CHI^2_GALAXY/CHI^2_STAR , CHI^2_QSO/CHI^2_STAR , and CHI^2_GALAXY/CHI^2_QSO . Their distributions are shown in Figure 4.14.

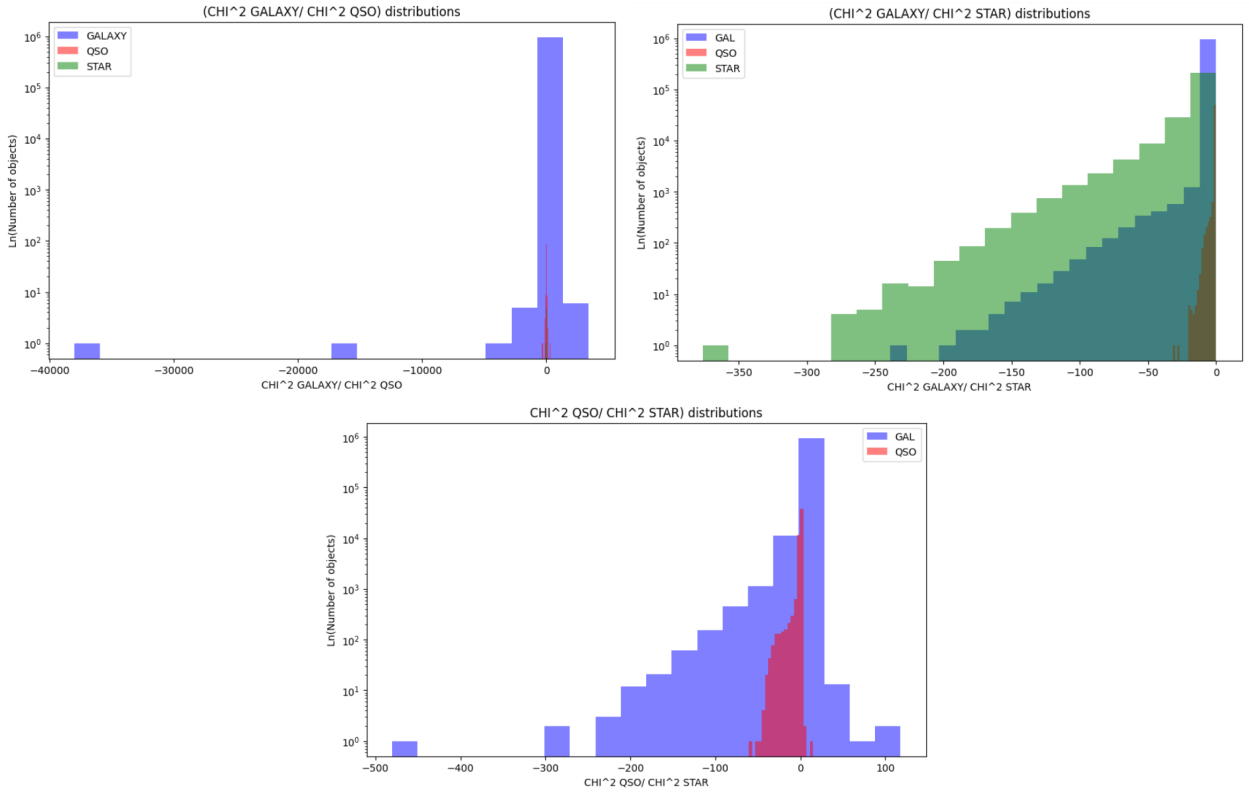

Figure 4.14: The distributions of each of the chi-square ratios values from the three templates used.

4.4.2 Flux inputs

The convolutional neural network, SAM, accepts as input the complete array of flux from the spectra. However, since the flux within the u band showed a relevant clue for differentiation between the spectral types, we also built a model that receives just the flux inside the u band as input.

One point that we have to consider when reading the flux input to train the network is the fact that this kind of models receive this entry as a tensor. Therefore, we have to reshape the original flux array as a tensor that looks like the input shown below.

```
array([[[-2.0244894 ],
        [-1.2554272 ],
        [ 0.14935206],
        ...,
        [ 0.1700719 ],
```

```
 [ 0.1937327 ],
 [ 0.05881049]],

[[-0.38240528],
 [-0.843131  ],
 [ 0.1980905 ],
 ...,
 [-0.20629807],
 [-0.38168576],
 [-0.18583867]],

...,

[[ 0.7531635 ],
 [ 0.35570222],
 [ 1.8457799 ],
 ...,
 [ 0.5689044 ],
 [ 0.8409228 ],
 [ 0.5581645 ]],], dtype=float32)
```

4.5 H2O models and relevant features

After completing the entire set of features for training the learning models, which totals 55 features, we assign one last random feature to each of the spectra. This is done so that when evaluating the relevance of all features in the learning process, those whose relevance is lower than that of the random feature must be discarded. In summary, we have 56 features for the first stage of training, including the random one.

The complete list of features comprises: *FLUX_U*, *FLUX_G*, *FLUX_R*, *FLUX_I*, *FLUX_Z*, *R-Z*, *U-R*, *G-Z*, *I-Z*, *U-G*, *(G-Z)/(R-Z)*, *R/U*, *G/Z*, *R/Z*, *I/Z*, *U/G*, *U/Z*, *CURV*, *MAD*, *F0*, *F1*, *F2*, *F3*, *F0/Z*, *F0/G*, *F0/R*, *F0/I*, *F0/F1*, *F0/F2*, *F0/F3*, *F1/U*, *F1/Z*, *F1/R*, *F1/I*, *F1/F2*, *F2/Z*, *F2/R*, *F2/I*, *F2/F3*, *F3/U*, *F3/Z*, *F3/R*, *F3/I*, *U/R T*, *U/G T*, *G/Z T*, *R/Z T*, *I/Z T*, *CHI²_GALAXY* , *CHI²_STAR* , *CHI²_QSO* , *CHI²_GALAXY/CHI²_QSO* , *CHI²_GALAXY/CHI²_STAR* , *CHI²_QSO/CHI²_STAR* , *ABBE*, *RANDOM*.

As expected, the dense neural network modeled by H2O showed significantly deficient metrics, classifying all elements as galaxies. This is because neural networks are not adept at interpreting tabular data, such as the features we are using.

On the other hand, these features were tested in a Random Forest (RF) model, whose metrics were quite promising and satisfactory for a first round of training. An overall-F1 score of 0.9797 was obtained, with an overall-accuracy of 0.9635.

Finally, the learning model determined by H2O as the best model was the gradient boosted machine (GBM), with metrics similar to those of the Random Forest. Here, an overall-F1 score of 0.9810 was achieved, with an overall-accuracy of 0.9659. However, LightGBM presents some optimization advantages over the GBM models, such as faster training speed, lower memory usage, and better handling of large datasets. So, we expect that using LightGBM instead of GBM, we can achieve better results.

Based on all this, we decided to use lightGBM, random forest, and the convolutional neural network SAM as classification models. Subsequently, tuning of these models was carried out.

4.6 Training step: Making a predictive model

4.6.1 Random forest and lightGBM models

Using the *sklearn* package, classification models for Random Forest and LightGBM have been constructed. These models are trained using the features mentioned in this chapter to determine which ones are truly useful for classifying spectral types in DESI data and how accurate these models are for this task.

For both models, we aim to tune the training parameters to obtain the best classification metrics with the least complexity possible. These parameters include: the number of estimators, which dictates the number of random trees used to build the model; maximum depth, which limits the maximum depth allowed for each decision tree in the ensemble; class

weight, which is useful for addressing class imbalance and in this case, we consider the values *balanced* and *None*; the learning rate (just applied for the lightGBM model), which limits the contribution of each tree to the final ensemble and whose values must be within the domain [0,1]; finally, bootstrap controls whether resampling is used during the training of each tree.

This training was carried out with 764,494 spectra, of which 578,889 are galaxies, 30,732 are quasars, and 154,873 are stars. These correspond to 60% of the total number of objects in the bright and dark catalogs. Then, the metrics used to measure the behavior of the models globally and individually by each one of the classes are the accuracy, recall, and F1-score.

4.6.2 SAM model

Similar to the two models mentioned earlier, SAM was trained with the same number of spectra, but using their flux arrays. In this model, the following parameters were tuned: the number of filters in each convolution, the size of the kernel or filter for each convolution, activation function, strides, pooling size, dropout rate, and number of neurons in the dense layers. Finally, the metrics used are also accuracy, recall, and F1-score.

Chapter 5

Analysis and results

5.1 Model performance results

5.1.1 Random forest model

Optimal parameters were found for this model, including 66 estimators, with a maximum depth of 40, without class weight, and without bootstrap.

With these hyperparameters, the model exhibited the best metrics and the lowest complexity in the validation process, which took the value of 0.9880 for overall accuracy. The metrics obtained for each individual class are presented in table 5.2. Additionally, after the procedure of eliminating redundant variables, the model was able to learn using 47 variables, with the 4 most relevant being:

Rank	Feature	Importance
1	CHI_GALAXYSTAR	0.092450
2	CHI_GALAXY	0.082828
3	U_R	0.074833
4	FLUX_R	0.065188

Table 5.1: Here are the 4 most relevant features that the Random Forest model learned from for the classification task. The first column displays the ranking of features from the most important, while the third column shows the importance. All the 47 features are shown in the Appendix.

Initially, fewer features would be required for training. However, by reducing the complexity of the random forest model, it utilizes a significant portion of the available information to learn as much as possible.

	Accuracy	Recall	F1-score
GALAXY	0.99	1.00	0.99
QSO	0.97	0.81	0.88
STAR	1.00	0.99	0.99

Table 5.2: Metrics per class obtained in the classification of galaxies, quasars, and stars using the Random Forest model.

Figure 5.1 shows the confusion matrix, where the model's good performance in the classification process of spectral types using spectrophotometric features can be observed.

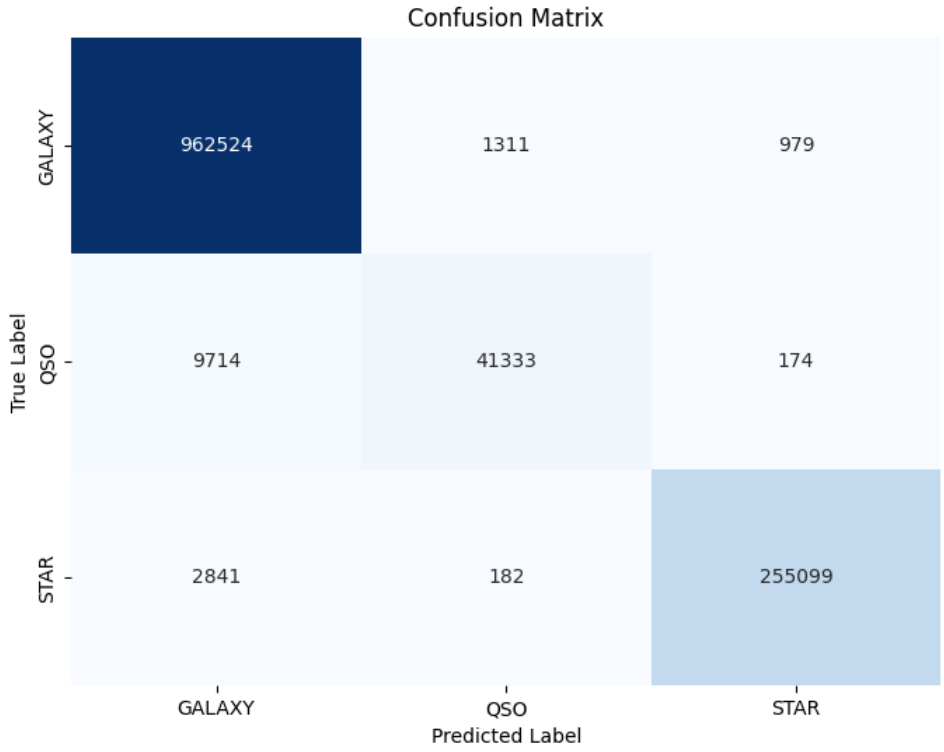

Figure 5.1: Confusion matrix of the Random Forest model. The objects on the diagonal represent those classes that were classified correctly.

Finally, probability histograms were created, where it is possible to appreciate with what certainty the model classifies or rejects an object as a galaxy, quasar, or star. In the ideal

case, each of these histograms should have a U shape, where a high frequency for probability values close to 0 implies that the model is confident that the object it is classifying does not correspond to that spectral class, while a high frequency in values close to 1 implies that it is confident that it corresponds to that class. If there are high frequencies at intermediate values, close to 0.5, it is bad, as it means that for those objects, the model is determining almost randomly whether it belongs to that class or not.

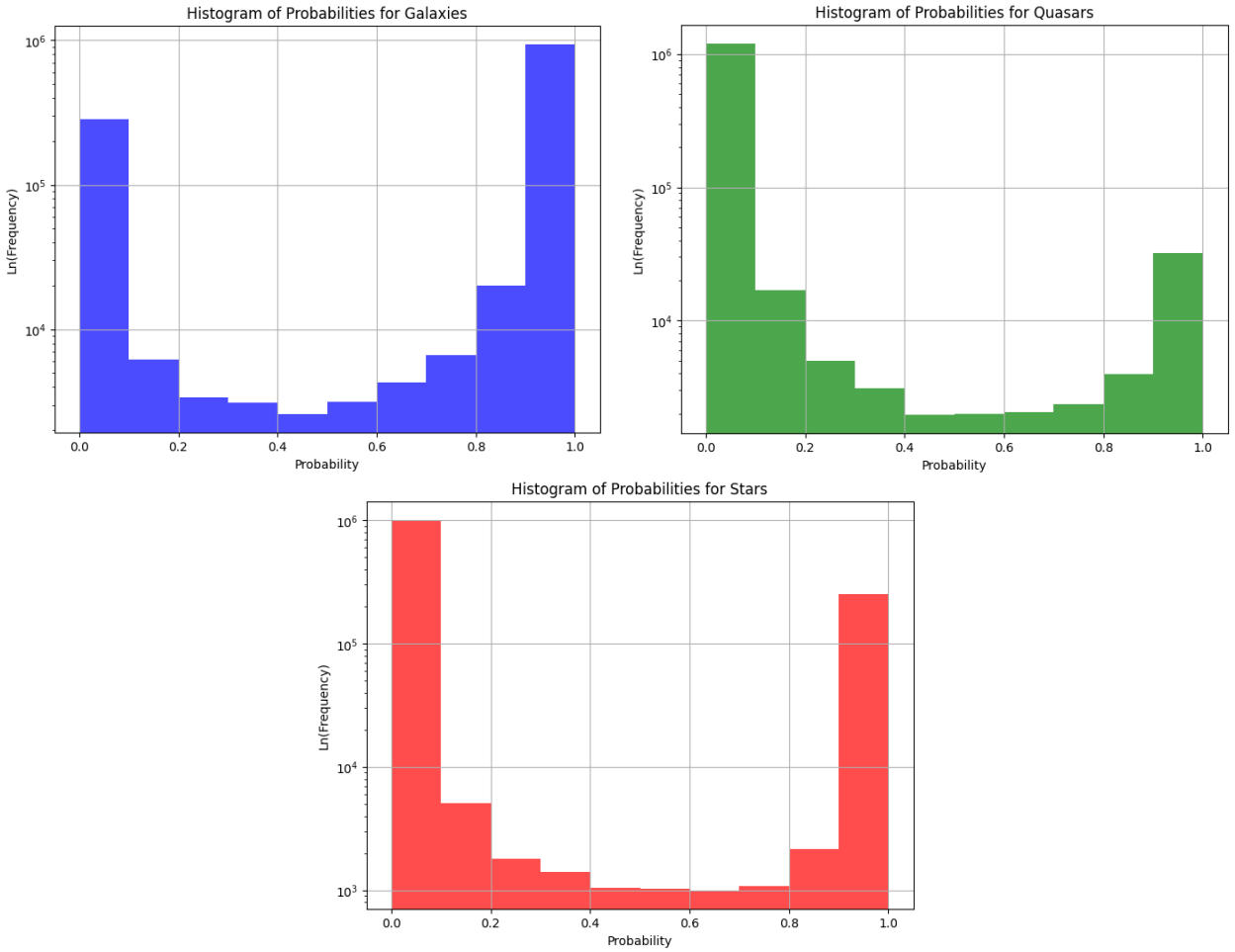

Figure 5.2: Probability histograms on the certainty when classifying objects from the DESI catalog as galaxies, quasars, or stars using the Random Forest model. The y-axis scale of each histogram is in the natural logarithm of the frequency.

From these histograms, it is clear that Random Forest classifies stars, quasars, and galaxies with high certainty. Therefore, this model is not guessing and provides us with reliable predictions.

The high accuracy values indicate that this model is correctly classifying each class and correctly discarding elements as galaxies, quasars, and stars. The recall is maximum for galaxies and stars because almost all of these objects have been classified correctly, but it is low for quasars because it is predicting some quasars as galaxies due to the problem of polarimetry.

5.1.2 LightGBM model

Similarly, optimal parameters were identified for the LightGBM model, encompassing 100 estimators, without establishing a maximum depth, a learning rate of 0.05, without class weight, and with bootstrap of 0.4.

Consistent with earlier findings, these hyperparameters resulted in the model showcasing the most favorable metrics during validation, with a value of 0.9755 for overall accuracy. Details regarding the metrics for each individual class are outlined in Table 5.4. Moreover, post-elimination of redundant variables, the model effectively learned from 31 variables. Concerning the top four, they are:

Rank	**Feature**	**Importance**
1	MAD	0.0651
2	ABBE	0.0574
3	CURV	0.0478
4	G_Z_R_Z	0.0447

Table 5.3: Here are the 4 most relevant features that the lightGBM model learned from for the classification task. The first column displays the rank, the second column shows the feature names, and the third column shows the importance. All the 31 features are shown in the Appendix.

	Accuracy	Recall	F1-score
GALAXY	0.97	0.99	0.98
QSO	0.89	0.61	0.72
STAR	0.99	0.98	0.98

Table 5.4: Metrics per class obtained in the classification of galaxies, quasars, and stars using the lightGBM model.

Figure 5.3 displays the confusion matrix, illustrating the model's adeptness in classifying spectral types utilizing spectrophotometric features.

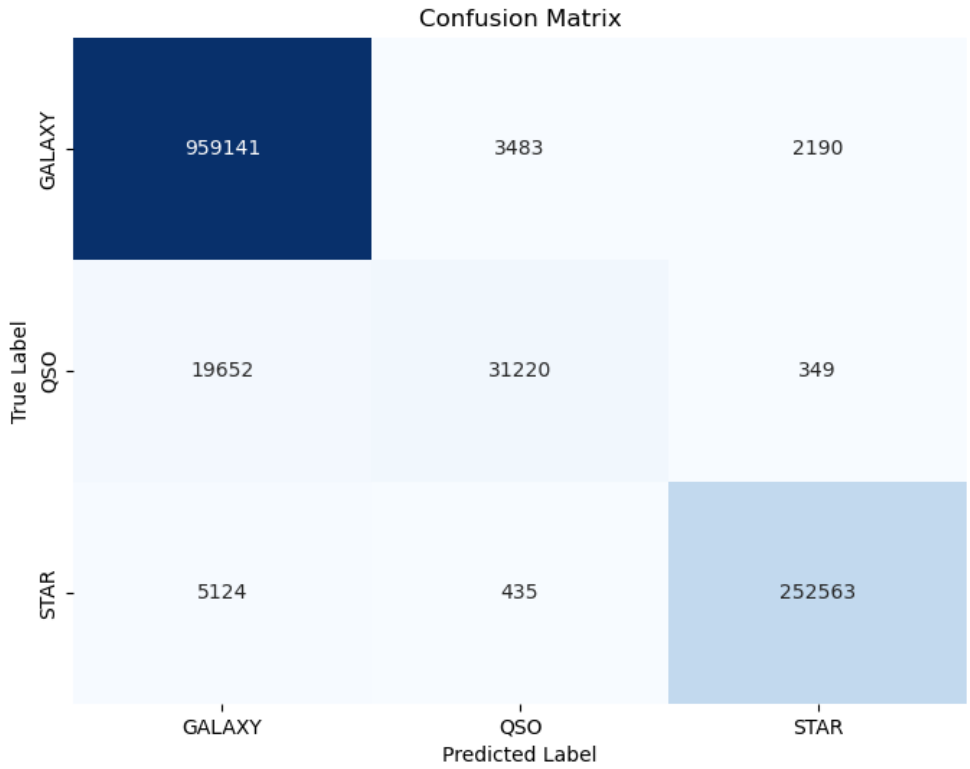

Figure 5.3: Confusion matrix of the lightGBM model. The objects on the diagonal represent those classes that were classified correctly.

Finally, as for the Random Forest, probability histograms were created and are shown in Figure 5.4.

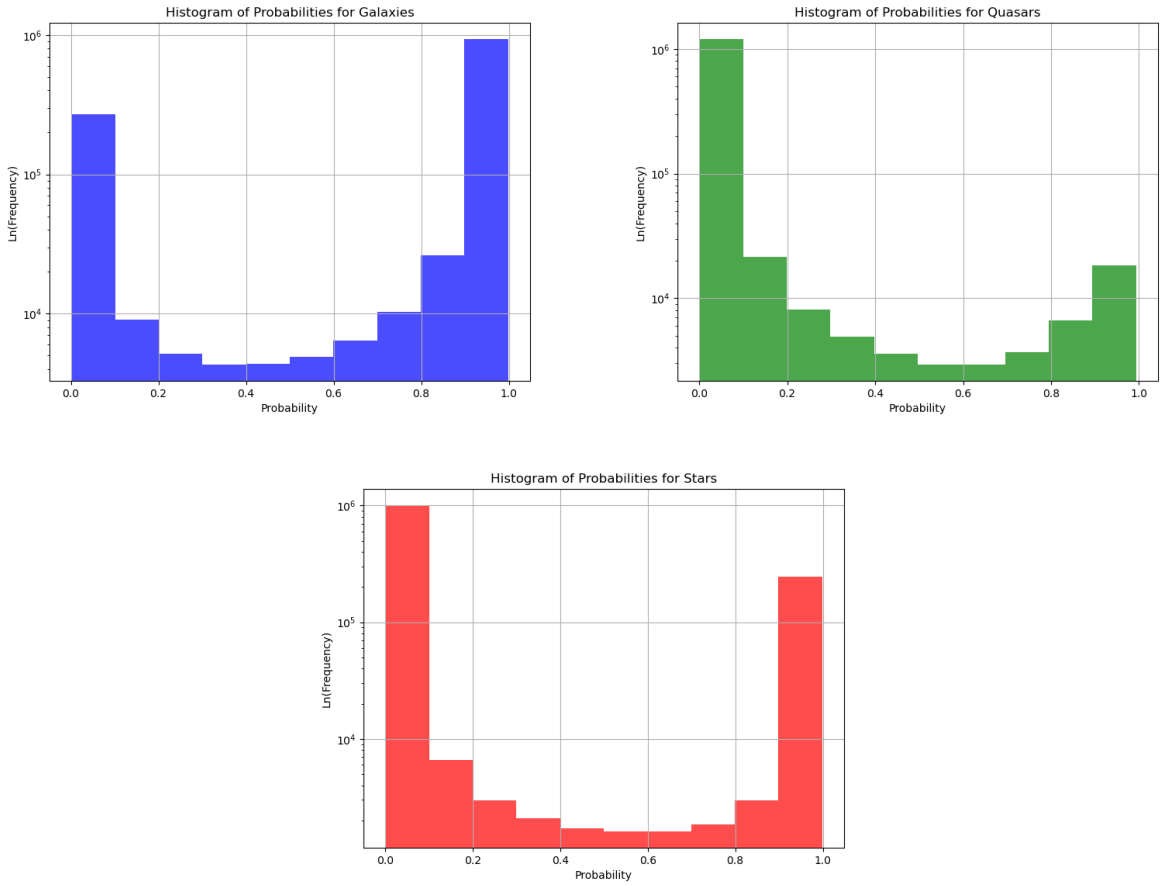

Figure 5.4: Probability histograms on the certainty when classifying objects from the DESI catalog as galaxies, quasars, or stars using the lightGBM model. The y-axis scale of each histogram is in the natural logarithm of the frequency.

The LightGBM model classifies stars, quasars, and galaxies with high certainty, and its predictions are reliable. We obtained high values for the metrics of galaxies and stars. On the other hand, the accuracy of the quasars is high, but the recall is low. This means that the model correctly discards many objects that are not quasars, but it also misclassifies many quasars as galaxies.

5.1.3 SAM model

Lastly, the optimal parameters for the Convolutional Neural Network (CNN) SAM were determined, including 64, 128, and 256 for the number of filters in the first, second, and

the last two convolutions, respectively. Also, 15 for the size of the kernel or filter for each convolution, *relu* as the activation function, 2 for the strides, 2 for the pooling size, 0.2 for the dropout rate, and 16 for the number of neurons in the dense layers.

Using this model, we achieved an overall accuracy of 0.9712. Detailed metrics for each individual class are provided in Table 5.5. Additionally, we found it preferable to utilize the full array of flux from the spectrum.

	Accuracy	Recall	F1-score
GALAXY	0.99	0.96	0.98
QSO	0.61	0.92	0.73
STAR	0.98	0.99	0.99

Table 5.5: Metrics per class obtained in the classification of galaxies, quasars, and stars using SAM.

Figure 5.5 illustrates the confusion matrix, demonstrating the model's improvement in classifying spectral types with respect to the previous version of SAM.

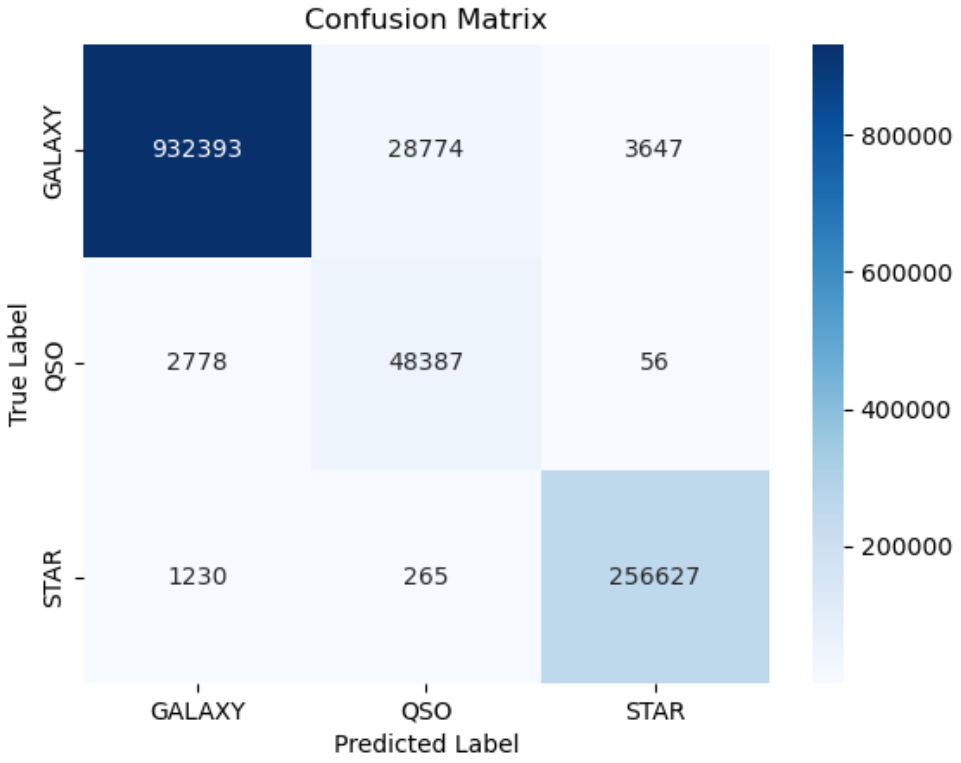

Figure 5.5: Confusion matrix of SAM. The objects on the diagonal represent correctly classified classes.

With respect to the histogram diagrams, we obtained slightly less certain decisions for galaxies and quasars with SAM compared to the other models. However, it also achieved high metrics in classifying these classes.

Although the QSO accuracy is the lowest of the three models, the recall is the highest. This means that SAM classifies most quasars as quasars but also misclassifies many galaxies as quasars.

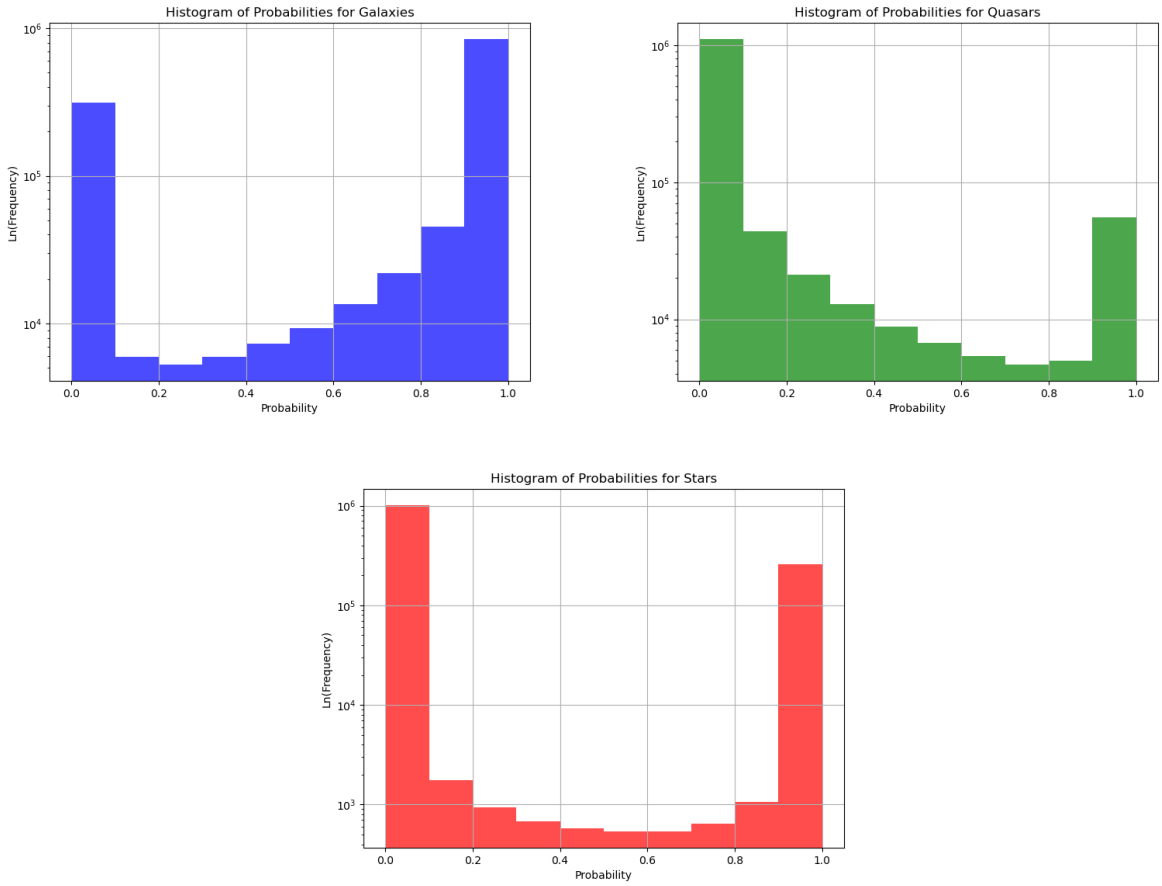

Figure 5.6: Probability histograms on the certainty when classifying objects from the DESI catalog as galaxies, quasars, or stars using the CNN SAM. The y-axis scale of each histogram is in the natural logarithm of the frequency.

5.1.4 Correlation matrix and relevant features

In an attempt to find a basis in the results obtained regarding which features the models deemed relevant for classification, the correlation between each of these 56 features was calculated using the Pearson correlation coefficient. These results were then depicted in the matrix shown in Figure 5.7.

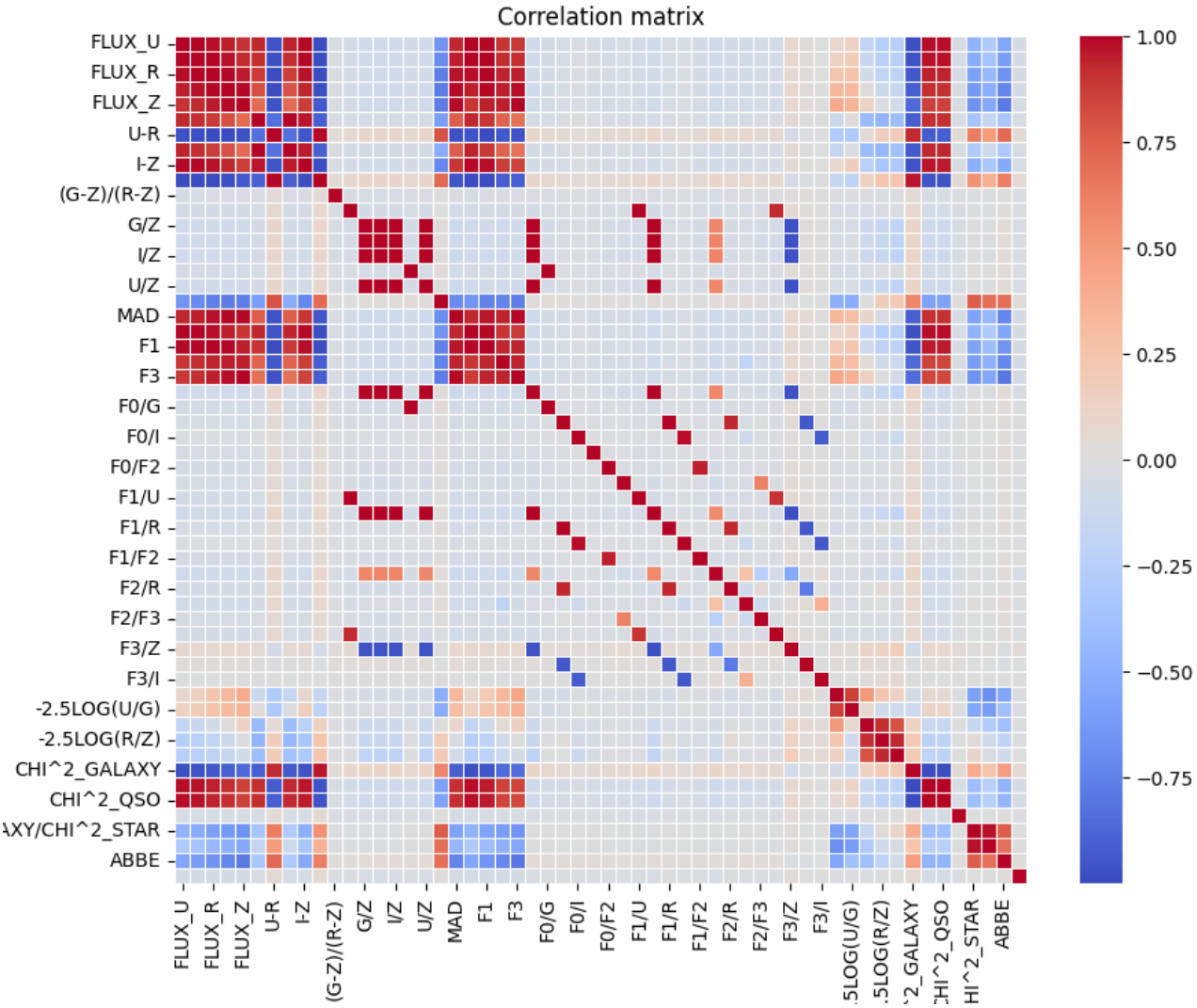

Figure 5.7: Correlation matrix between each of the 56 features used in the training process of the classification models. Each row and column corresponds to a feature. However, the axes do not display the names of the 55 features to avoid image saturation, but the order in which they appear is the same as in the list presented in section 4.5.

In this matrix, a correlation value of 1 between two features means that both change proportionally to each other. On the other hand, if the value is -1, their relative change is inversely proportional. While a value of 0 indicates no correlation between both features.

The correlation matrix reveals that the relevant features chosen by the models are those that are highly correlated with the ugriz flux features, color indices, synthetic fluxes, chi-

squared values and their ratios, and the curvature, MAD, and Abbe parameters, as well as showing considerable correlation with color temperatures. Whereas, nearly all features corresponding to the ratios between these quantities, which sought to exhibit a clearer rate of change between them, are discarded due to their almost negligible correlation with the relevant features. Finally, as a point worth noting, the random feature does not correlate with any other, as expected.

Once we are left only with the features that meet the requirement of being relevant, it entirely depends on the model to determine which of these are redundant with others, and thus not consider them relevant since they can learn the same from other features. However, as a general trend, all relevant features have a high correlation among them.

5.2 Constructing a catalogue

Although each of the three classifiers presents high values of accuracy, recall, and F1 score for galaxies and stars, the Random Forest and lightGBM models have high accuracy and low recall, meaning that these models correctly discard many objects as quasars but also discard many real quasars. On the other hand, SAM has a high recall but low accuracy, meaning that it correctly classifies almost all quasars but also misclassifies many galaxies as quasars.

Additionally, the probability histograms make evident the fact that these models have high certainty when labeling or discarding an object as a galaxy, quasar, or star. Therefore, we decided to use a strategy for constructing the new catalog that labels the objects within the bright and dark catalogs of DESI as galaxies if at least two of the three implemented classifiers agreed that they corresponded to that class. Similarly, quasars and stars were labeled. This way, the three models can balance the final decision on the label of each object, reducing the misclassification of quasars and galaxies by compensating for the low recall of some models and the low accuracy of others.

Those objects that received a different prediction from each of the three models are

considered objects that require review and therefore their label has been changed to *EXAMINE* instead of *GALAXY*, *STAR*, or *QSO*. In this way, we have made the classification of objects using the spectrophotometric features calculated in Chapter 4, thus avoiding the problems of Seyfert line types and the problematic polarimetry of observed quasars described in Chapter 1. This new catalog is called SPAß, referring to the acronym SPASS derived from the name spectrophotometric assembly classifiers.

In Table 5.6, the percentage of coincidences and differences in the determination of the label of objects from the bright and dark catalogs of DESI are shown.

	Percentage of coincidences	Percentage of differences
GALAXY	99.57%	0.43%
QSO	78.39%	21.61%
STAR	98.70%	1.30%

Table 5.6: Percentage of coincidences and differences in the labels of galaxies, quasars, and stars between the DESI's catalogue and the SPAß catalogue.

Subsequently, we generated a fits file containing 10 tables, each corresponding to one of the 10 files used to store the DESI spectra referenced in Chapter 4. However, within each of these tables, the spectral classes determined by our strategy are listed. This fits file is available in our repository under the name *SPECTYPE_SPASS.fits*[1].

5.3 Mapping the Universe

Finally, we have approached the distribution of stars, quasars, and galaxies in our observable universe according to the spectral classes within the SPAß catalog along with their corresponding coordinates taken from the DESI catalog. The distributions are illustrated in Figures 5.8 and 5.9.

[1] Virtual GitHub repository: https://github.com/DanielFajardo1/Mapping-the-Universe.git

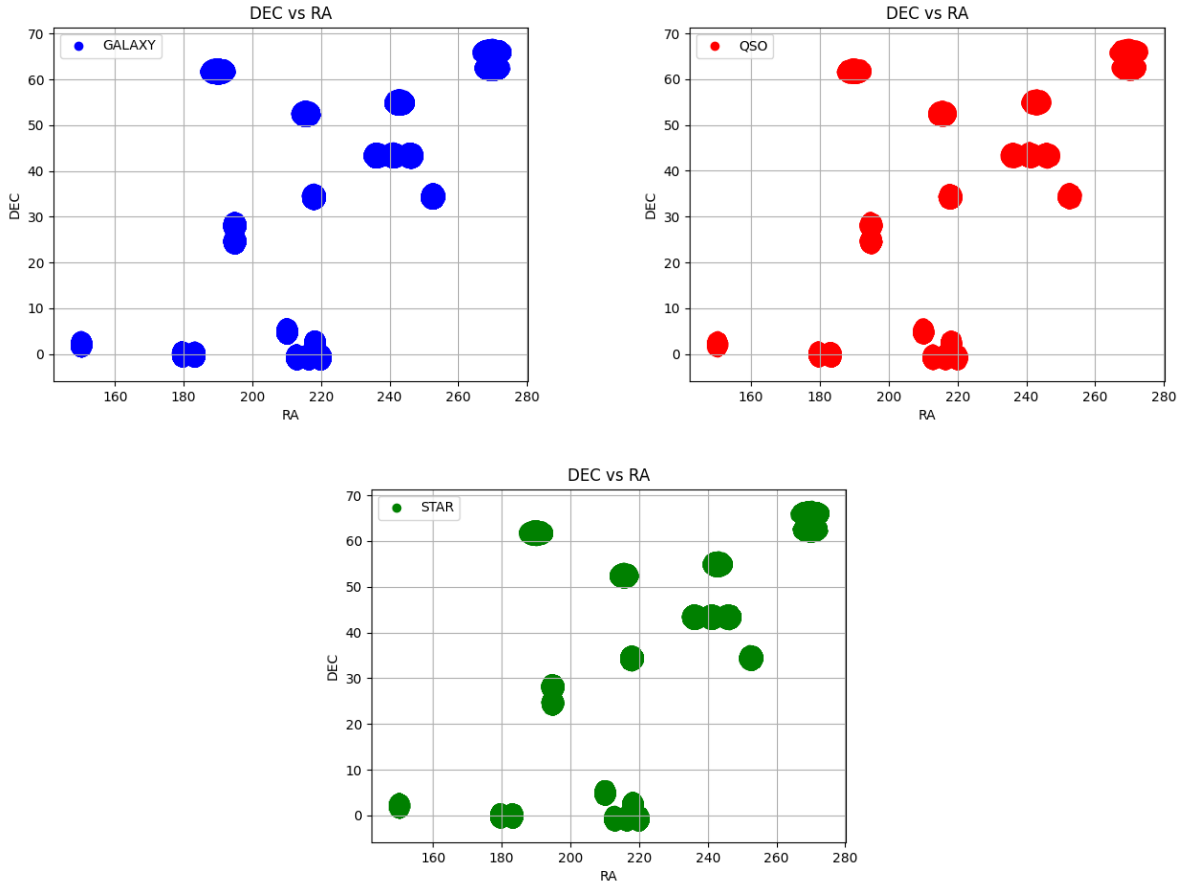

Figure 5.8: Distribution of galaxies, quasars, and stars according to the spectral classes and coordinates, declination and right ascension, from the SPAß and DESI catalogs.

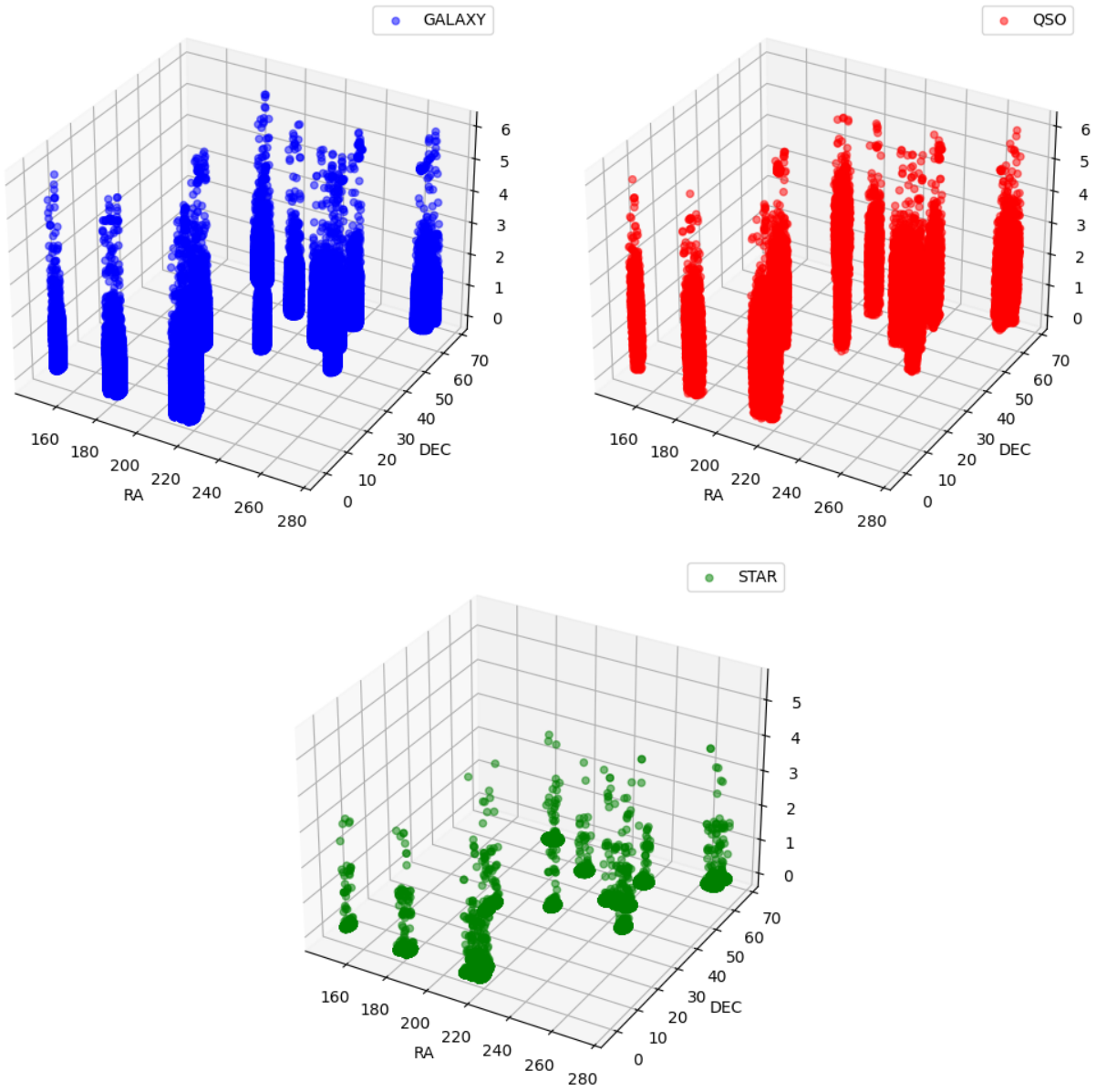

Figure 5.9: Distribution of galaxies, quasars, and stars according to the spectral classes and 3D-coordinates from the SPAß and DESI catalogs.

These distributions match the characteristics specific to each of the spectral classes. In the case of stars, we see that they saturate at low redshift values, because they are stars from our galaxy and, therefore, as expected, are not as distant as the quasars or galaxies observed. However, those stars found with high redshift values allowed the DESI collaboration to confirm the migration of stars into the Andromeda galaxy (Myers et al. 2023).

On the other hand, galaxies have a distribution that extends mainly up to redshift values of 3, but there are also galaxies that exceed this threshold, though they are not a significant part of the total population.

Quasars, meanwhile, span the widest range of values, from very small values, where ambiguity arises with the morphologies of these and those of stars and galaxies, to extremely large values, with a large population of quasars reaching redshift values close to 6. These high redshift quasars are the best studied quasars by the DESI collaboration, those with z>2.1.

Additionally, the 20 observation windows chosen by DESI to maximize the completeness of each class in their catalogs are evident (Chaussidon et al. 2023). That is, these windows maximize the number of quasars and galaxies as deep space observations are made. For this reason, the three spectral classes are seen within the same observation fields and, in the case of the 3D image, pillars that allow us to get an idea of the distribution of objects in our universe.

Chapter 6

Conclusions

At the end of this study, the utility of spectrophotometric features for the classification exercise using spectra measured by the DESI collaboration has become clear.

Metrics such as accuracy, recall, and F1 score were obtained, which were high for classes such as stars and galaxies. A strategy was implemented to deal with the low performance of some models in certain metrics, compensating with the good performance of others. This strategy involved classifying an object as a galaxy, quasar, or star if at least two out of three models classified it as such. This approach led to the creation of a new catalog of objects, labeled based on this spectrophotometric treatment.

This catalog, SPAß, showed a significant difference in the number of objects classified as quasars compared to those classified within the DESI catalog. The differences arise from the strategies used for the determination of this class.

On the other hand, the selected models used to maximize classification metrics and reduce complexity were Random Forest, lightGBM as an application of the initial gradient boosted machine model, and the convolutional neural network SAM. It was found that the first two models discard a high quantity of true quasars as quasars in order not to contaminate the total sample but decreasing the completeness of the catalog, while SAM, conversely, accepts some galaxies and stars as quasars to avoid reducing the completeness of the final sample but

reducing accuracy.

Then, using the spectral classes within the SPAß catalog, the distribution of galaxies, quasars, and stars in our Universe was visualized according to the measurements reported by the DESI contribution. In these distributions, the 20 selected observation fields were evident, aiming to maximize the number of galaxies and quasars in the sample of measured objects. Additionally, the results perfectly coincide with what is expected for each class, where stars have the smallest redshift values, followed by galaxies, and then by quasars.

As future work, it is suggested to perform a statistical analysis on the SPAß catalog in order to corroborate the expected population results of stars, quasars, and galaxies according to the literature regarding the redshift distribution. It is also proposed to predict the redshift of all objects in the dark and bright catalogs using the spectrophotometric features found in this work, to address the issues that arise from using only templates.

Additionally, the tools developed in this work have great potential to contribute to the DESI collaboration. Therefore, it would be interesting to use them in noisier catalogs, such as the backup catalog, where some tools like line classifiers lose precision, to carry out the classification of objects and the calculation of other physical properties such as redshift.

Finally, using the quasars and all the information within the SPAß and DESI catalogs, it is suggested to contribute to the investigation of quasar by utilizing some of the features presented, such as color temperature, in order to understand the detailed properties of the different types of spectra we can receive from a quasar depending on the angle at which it is observed. This would determine their morphology, their emission lines, and verify theoretical hypotheses about their structures composed of a torus and active nuclei. Subsequently, an analysis would be conducted to determine the subtypes of quasars.

Bibliography

Adam, G. 2022, AJ, 163, 11

Adame, A. G. et al. 2023, arXiv preprint: [2306.06308]

Chaussidon, E. et al. 2023, ApJ, 944, 107

Cárdenas, C. & Fajardo, D. 2023, theoretical project, Universidad de los Andes

Farr, J. et al. 2020, Journal of Cosmology and Astroparticle Physics, 2020, 15

Gallagher, S. C. et al. 2015, RAS, 451, 2991

Ginsburg, A. et al. 2022, ApJ, 163, 291

Karttunen, H. et al. 2007, Fundamental Astronomy, Springer

Li, R. et al. 2019, MNRAS, 482, 313

Li, Z. et al. 2022, MNRAS, 517, 4875

Myers, A. D. et al. 2023, Astron. J, 165, 50

Ness, M. et al. 2015, ApJ, 808, 16

Ostlie, A. & Carroll, W. 1996, Modern Stellar Astrophysics, Addison-Wesley Publishing Company

Robitaille, T. P. et al. 2013, A&A, 558, A33

Saavedra, J. 2019, undergraduate thesis, Universidad de los Andes, 59

Wang, B. et al. 2022, ApJS, 259, 28

Zakamska, N. & Alexandroff, R. 2023, arXiv preprint: [2306.06303]

Appendices

Appendix A

Photometric band fluxes

Similar behavior was observed among the distributions of all fluxes across the ugriz bands. Here are the distributions and box plots of these features:

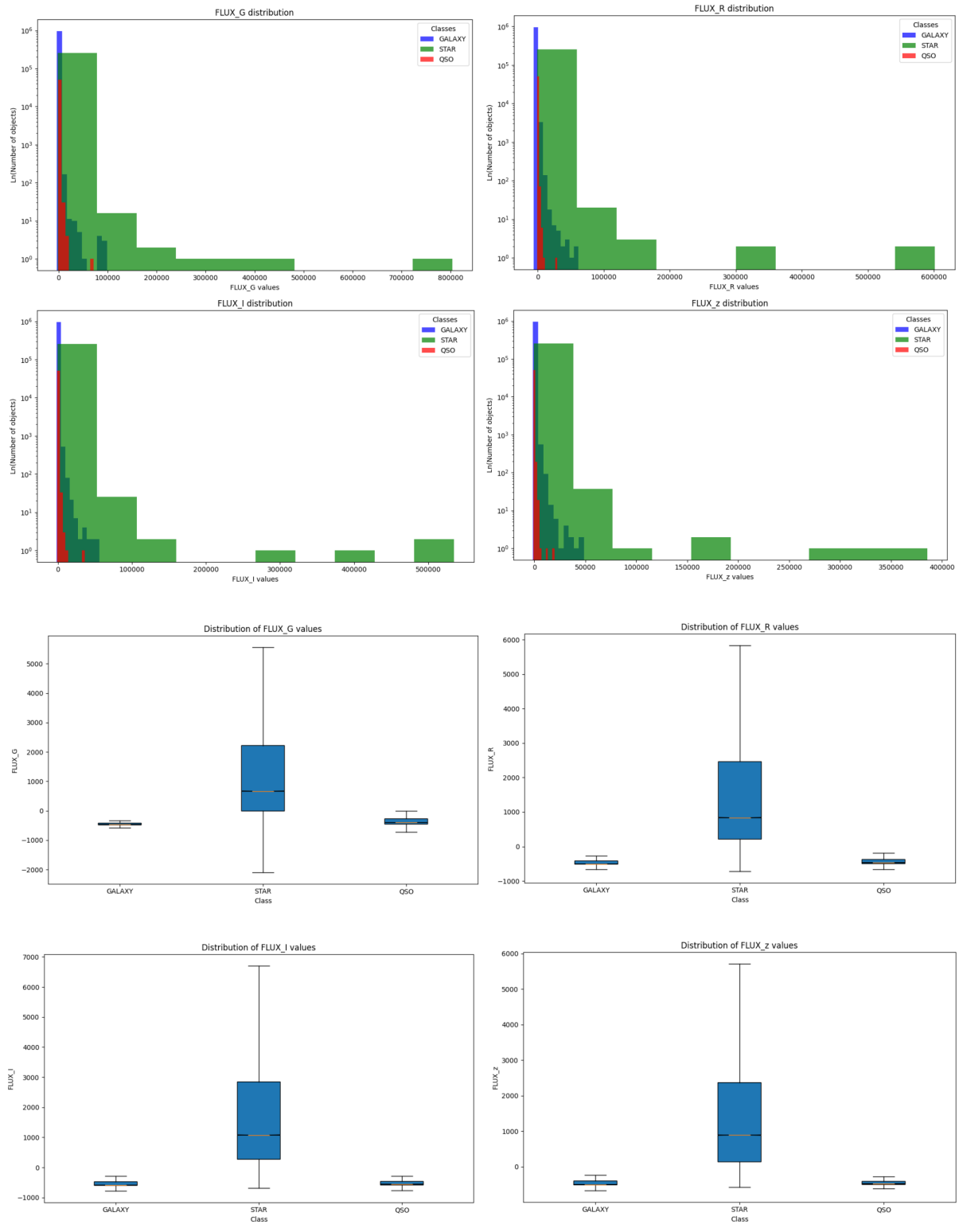

Figure A.1: Distributions of flux values for the ugriz bands.

Appendix B

Band fluxes slopes

The majority of the slopes between flux values under the photometric bands do not show significant differentiation among the various spectral types or are redundant. For this reason, after conducting correlation analysis and assessing the relevance of these features, all were discarded. All the sets of ratios that we initially considered, including the ratios from other features, are: G/R, R/Z, R/U, R/I, U/G, U/I, U/Z, G/I, G/Z, I/Z, (U-R)/(G-Z),(U-R)/(R-Z), (U-R)/(I-Z), (U-R)/(U-G), (G-Z)/(R-Z), (G-Z)/(I-Z), (G-Z)/(U-G), (R-Z)/(I-Z), (R-Z)/(U-G), (I-Z)/(U-G), F0/U, F0/G, F0/R, F0/I, F0/Z, F0/F1, F0/F2, F0/F3, F1/U, F1/G, F1/R, F1/I, F1/Z, F1/F2, F1/F3, F2/U, F2/G, F2/R, F2/I, F2/Z, F2/F3, F3/U, F3/G, F3/R, F3/I, F3/Z, (U/R T)/(G/Z T), (U/R T)/(R/Z T), (U/R T)/(I/Z T), (U/R T)/(U/G T), (G/Z T)/(R/Z T), (G/Z T)/(I/Z T), (G/Z T)/(U/G T), (R/Z T)/(I/Z T), (R/Z T)/(U/G T), (I/Z T)/(U/G T), (U-R)/U, (U-R)/G, (U-R)/R, (U-R)/I, (U-R)/Z, (U-G)/U, (U-G)/G, (U-G)/R, (U-G)/I, (U-G)/Z, (G-Z)/U, (G-Z)/G, (G-Z)/R, (G-Z)/I, (G-Z)/Z, (R-Z)/U, (R-Z)/G, (R-Z)/R, (R-Z)/I, (R-Z)/Z, (I-Z)/U, (I-Z)/G, (I-Z)/R, (I-Z)/I, (I-Z)/Z. These are the scatter plots of some of these relationships:

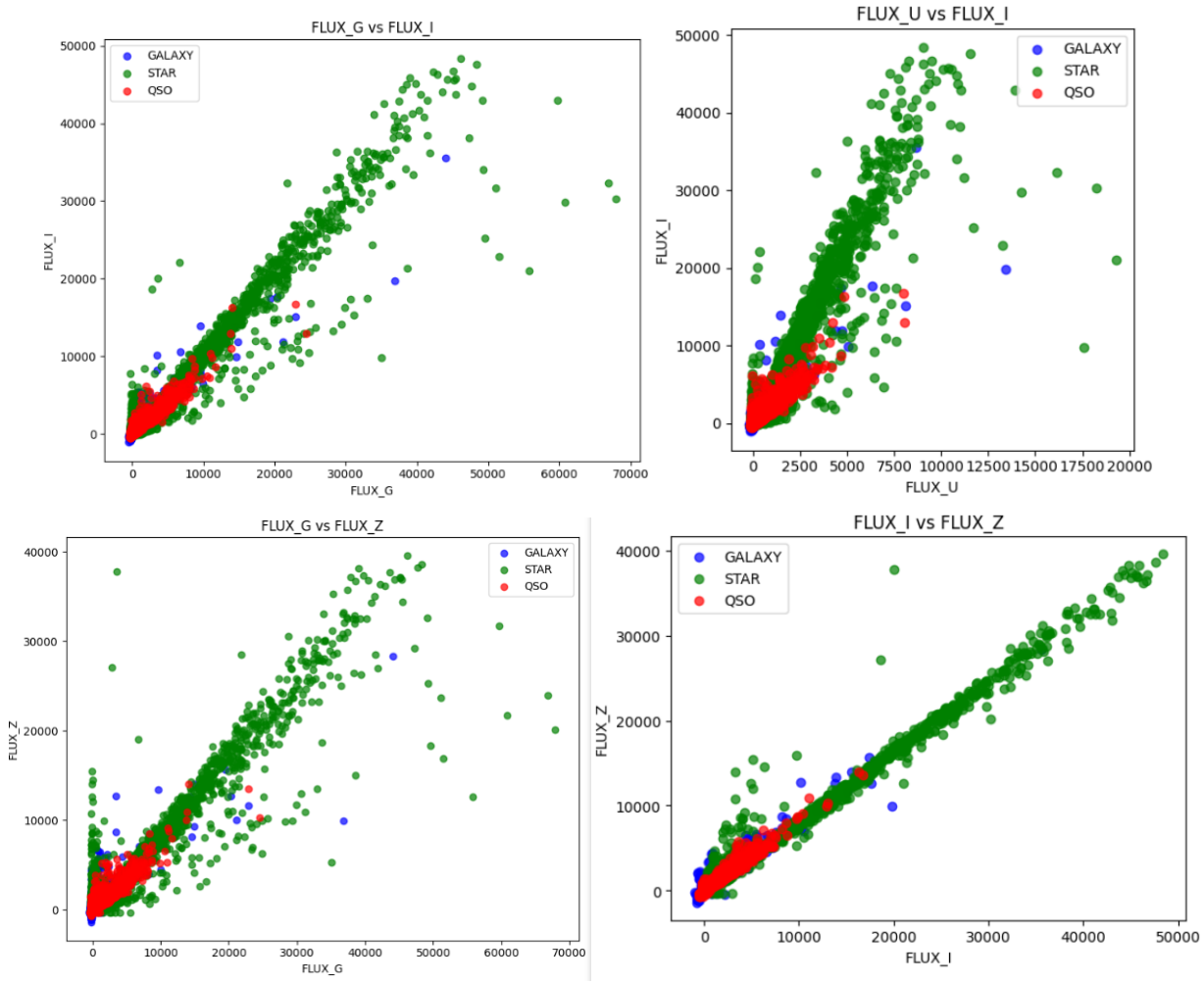

Figure B.1: Scatter plots of some bands against others. Here, one can appreciate the slopes at which the objects lie.

Appendix C

χ^2 Templates

As expected from the values obtained for the chi-square homogeneity tests, each of the spectra presents a value much closer to zero with respect to the template corresponding to the mean flux value of its class, compared to the values obtained from the chi-square test using the other templates. These are the box and whisker plots that show this behavior:

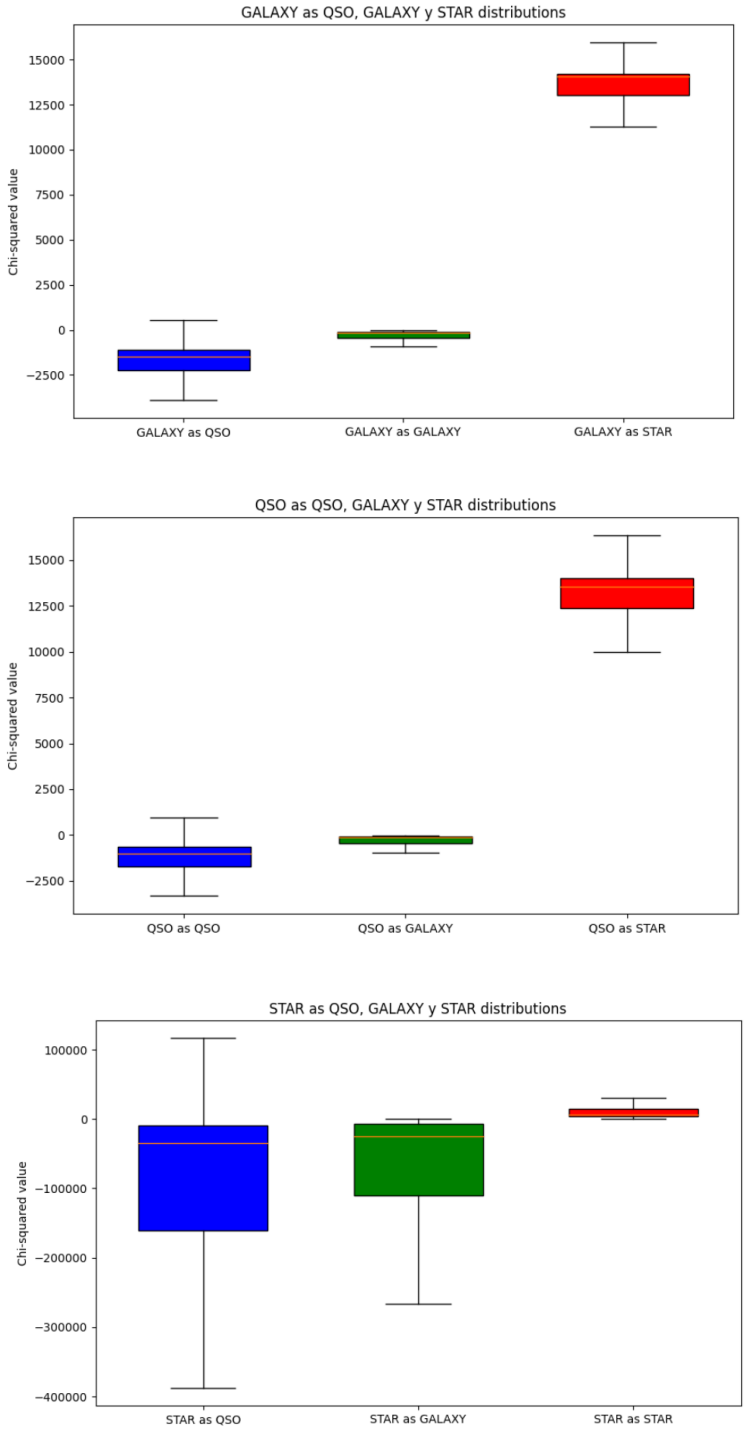

Figure C.1: Box plots exhibiting the expected behavior of similarity of the spectra with respect to the template using the mean flux value of each class. The top figure displays how much galaxies differ from the template of quasars, galaxies, and stars. Meanwhile, the middle figure shows how much quasars differ from the same three templates, and the bottom figure illustrates the behavior for stars.

Appendix D

Relevant Features

D.1 Random Forest

Here are all the 47 variables considered relevant by the random forest model for learning and classifying the spectra:

Rank	Feature	Importance
1	CHI_GALAXYSTAR	0.092450
2	CHI_GALAXY	0.082828
3	U_R	0.074833
4	FLUX_R	0.065188
5	FLUX_I	0.063793
6	ABBE	0.057522
7	F1	0.043778
8	FLUX_G	0.041758
9	MAD	0.035459
10	U_G	0.034347
11	FLUX_Z	0.032533

Table D.1 – Continued from the previous page

Rank	Feature	Importance
12	CHI_QSOSTAR	0.029410
13	F2	0.026917
14	F3	0.024846
15	G_Z	0.023804
16	FLUX_U	0.023050
17	I_Z	0.019536
18	R_Z	0.018843
19	G_Z_R_Z	0.013044
20	UZ	0.012307
21	CHI_GALAXYQSO	0.009887
22	RU	0.009759
23	UG	0.009520
24	CHI_QSO	0.008968
25	CHI_STAR	0.008819
26	GZ	0.008052
27	T_GZ	0.007894
28	T_UG	0.007734
29	_3U	0.007479
30	T_IZ	0.006782
31	F0	0.006741
32	T_RZ	0.006434
33	T_UR	0.005982
34	_3I	0.005906
35	_03	0.005817
36	_0R	0.005679

Table D.1 – Continued from the previous page

Rank	Feature	Importance
37	_3R	0.005655
38	_0I	0.005271
39	IZ	0.005251
40	RZ	0.005121
41	_0Z	0.004880
42	_0G	0.004280
43	_01	0.003844
44	_3Z	0.003532
45	_1U	0.003177
46	_02	0.002705
47	_2R	0.002397

Table D.1: All relevant features for the Random Forest model. The first column displays the ranking of features from the most important, while the third column shows the importance.

D.2 lightGBM

Here are all the 31 variables considered relevant by the lightGBM model for learning and classifying the spectra:

Rank	Feature	Importance
1	MAD	0.0651
2	ABBE	0.0574
3	CURV	0.0478
4	G_Z_R_Z	0.0447

Table D.2 – Continued from the previous page

Rank	Feature	Importance
5	UG	0.0411
6	FLUX_U	0.0406
7	_3Z	0.0391
8	R_Z	0.0391
9	U_G	0.0379
10	CHI_GALAXY	0.0364
11	T_UG	0.0362
12	I_Z	0.0322
13	_0G	0.0288
14	U_R	0.0280
15	FLUX_Z	0.0258
16	F0	0.0251
17	G_Z	0.0226
18	CHI_STAR	0.0212
19	FLUX_R	0.0190
20	CHI_GALAXYSTAR	0.0189
21	_0R	0.0157
22	_3I	0.0155
23	F3	0.0153
24	IZ	0.0152
25	FLUX_I	0.0143
26	UZ	0.0139
27	_3R	0.0134
28	FLUX_G	0.0130
29	_3U	0.0127

Table D.2 – Continued from the previous page

Rank	**Feature**	**Importance**
30	T_IZ	0.0121
31	_0I	0.0121

Table D.2: All relevant features for the lightGBM model. The first column displays the rank, the second column shows the feature names, and the third column shows the importance.

Appendix E

Advisor's signature

Figure E.1: Advisor's signature