Unveiling the physical conditions for Lyman-α photons escape using Neural Network Activations

Bruno Cerqueira
Master in Astronomy and Astrophysics
Department of Physics and Astronomy
Faculty of Sciences of the University of Porto
2025

Unveiling the physical conditions for Lyman- α photons escape using Neural Network Activations

Bruno Cerqueira

Dissertation carried out as part of the Master in Astronomy and Astrophysics

Department of Physics and Astronomy

2025

Supervisor
Dr. Ana Paulino-Afonso,
Researcher, Instituto de Astrofísica e Ciências do Espaço,
Universidade do Porto
Co-supervisor
Dr. José Fonseca,
Researcher, Instituto de Astrofísica e Ciências do Espaço,
Universidade do Porto

fct Fundação para a Ciência e a Tecnologia

ia instituto de astrofísica e ciências do espaço

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Dra. Ana Paulino-Afonso, Dr. José Fonseca, and Dr. Andrew Humphrey for their unwavering support and invaluable guidance throughout this journey. A special thanks to MSc. Afonso Vale, whose technical assistance and continuous encouragement were truly appreciated.

I am also immensely thankful to my parents, Jose Cicero and Cleonildes and my brother, Daniel, for giving me the opportunity to pursue this path and for their unwavering emotional support along the way. A special mention goes to my cat, Banguela, and my dog, Orion, who probably know better than anyone how many hours I spent researching instead of brushing them.

My sincere thanks also go to the IA/CAUP team for their insightful discussions and collaborative environment, and to the professors and staff at FCUP, whose dedication helped shape a rigorous and inspiring academic experience.

I, Bruno Barbosa Cerqueira, acknowledge funding by Fundação para a Ciência e a Tecnologia (FCT) through the research grants UIDB/04434/2020 and UIDP/04434/2020, and in the form of an exploratory project with the EXPL/FIS-AST/1085/2021 reference (PI: Paulino-Afonso).

This research used COSMOS2020: A panchromatic view of the Universe to z - 10 from two complementary catalogues, Weaver et al. (2022). Based on observations collected at the European Southern Observatory under ESO programme ID 179.A-2005 and on data products produced by CALET and the Cambridge Astronomy Survey Unit on behalf of the UltraVISTA consortium.

HETDEX is led by the University of Texas at Austin McDonald Observatory and Department of Astronomy with participation from the Ludwig-Maximilians-Universität München, Max-Planck-Institut für Extraterrestrische Physik (MPE), Leibniz-Institut für Astrophysik Potsdam (AIP), Texas A&M University, Pennsylvania State University, Institut für Astrophysik Göttingen, The University of Oxford, Max-Planck-Institut für Astrophysik (MPA), The University of Tokyo and Missouri University of Science and Technology.

Observations for HETDEX were obtained with the Hobby-Eberly Telescope (HET),

which is a joint project of the University of Texas at Austin, the Pennsylvania State University, Ludwig-Maximilians-Universität München and Georg-August-Universität Göttingen. The HET is named in honor of its principal benefactors, William P. Hobby and Robert E. Eberly. The Visible Integral-field Replicable Unit Spectrograph (VIRUS) was used for HETDEX observations. VIRUS is a joint project of the University of Texas at Austin, Leibniz-Institut für Astrophysik Potsdam (AIP), Texas A&M University, Max-Planck-Institut fürExtraterrestrische Physik (MPE), Ludwig-Maximilians-Universität München, Pennsylvania State University, Institut für Astrophysik Göttingen, University of Oxford, and the Max-Planck-Institut fur Astrophysik (MPA).

The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high performance computing, visualization, and storage resources that have contributed to the research results reported within this paper. URL: http://www.tacc.utexas.edu

Funding for HETDEX has been provided by the partner institutions, the National Science Foundation, the State of Texas, the US Air Force, and by generous support from private individuals and foundations.

Finally, I extend my appreciation to the SC4K team [1] for making their catalog publicly available, an essential resource that made this work possible.

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto
Departamento de Física e Astronomia

MSc. Astronomy and Astrophysics

Unveiling the physical conditions for Lyman-α photons escape using Neural Network Activations

by Bruno CERQUEIRA

Understanding the conditions that allow Lyman-α photons to escape from galaxies is essential for studying galaxy evolution and the early Universe. Traditional LAE (Lyman-α emitter) selection methods, such as narrow-band imaging or spectroscopic follow-up, limit the efficiency and scalability of identifying these objects in large photometric surveys. This dissertation investigates whether deep learning models, specifically convolutional neural networks (CNNs), can identify LAEs and estimate their emission properties using only broadband imaging. RGB composite images were built from broadband photometry using sources from the SC4K and COSMOS2020 catalogs. A CNN was trained to classify LAEs and nLAEs, while separate regression models were developed to estimate redshift, Lyman-α luminosity, and equivalent width. The classifier achieved an overall accuracy of $\sim$**75.84**%, with a F1-score of **75.51**% across perturbed datasets. Cross-matching with HETDEX sources showed the model maintained a high mean precision of **86.41**%. The regression models achieved an average absolute error (MAE) of **0.032** for redshift, for log equivalent width, and for Lyman-α luminosity. Saliency maps and perturbation analysis revealed that the models primarily focus on compact, central regions of the sources, consistent with known properties of LAEs. This suggests that CNNs can learn physical features relevant to Lyman-α escape. This work shows that deep learning can complement traditional LAE selection, enabling scalable analysis of photometric data and supporting future applications in surveys like *Euclid* and *MOONS*.

Keywords: galaxy, Lyman-α emitters, Lyman-α escape, broadband photometry, convolutional neural networks, high redshift, saliency maps, SC4K

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto
Departamento de Física e Astronomia

Mestrado Astronomia e Astrofísica

Desvendando as condições físicas para a fuga de fótons Lyman-α usando Ativações de Redes Neurais

por Bruno CERQUEIRA

Compreender as condições que permitem a fuga de fótons de Lyman-α das galáxias é essencial para o estudo da evolução galáctica e do Universo primordial. Métodos tradicionais de seleção de LAEs (emissores de Lyman-α), como imageamento com filtros de banda estreita ou acompanhamento espectroscópico, limitam a eficiência e a escalabilidade na identificação desses objetos em grandes levantamentos fotométricos. Esta dissertação investiga se modelos de aprendizado profundo, especificamente redes neurais convolucionais (CNNs), podem identificar LAEs e estimar suas propriedades de emissão utilizando apenas imagens em banda larga. Imagens RGB compostas foram construídas a partir de fotometria em banda larga usando fontes dos catálogos SC4K e COSMOS2020. Uma CNN foi treinada para classificar LAEs e nLAEs, enquanto modelos de regressão independentes foram desenvolvidos para estimar redshift, luminosidade de Lyman-α e largura de linha equivalente. O classificador alcançou uma acurácia global de aproximadamente **75.84**%, com uma mediana de F1-score de **75.51**% entre os diferentes conjuntos de dados perturbados. O cruzamento com fontes do catálogo HETDEX demonstrou que o modelo manteve uma precisão média elevada de **86.41**%. Os modelos de regressão obtiveram um erro absoluto médio (MAE) de **0.032** para redshift, log da largura de linha equivalente, e luminosidade de Lyman-α. Mapas de saliência e análises de perturbação revelaram que os modelos concentram-se principalmente nas regiões centrais e compactas das fontes, em concordância com as propriedades conhecidas de LAEs. Isso sugere que as CNNs conseguem aprender características físicas relevantes para a fuga de fótons de Lyman-α. Este trabalho mostra que o uso de deep learning pode complementar a seleção

tradicional de LAEs, permitindo uma análise escalável de dados fotométricos e apoiando aplicações futuras em levantamentos como o *Euclid* e o *MOONS*.

Palavras-chave: galáxias, emissores de Lyman-α, escape de Lyman-α, fotometria de bandas largas, redes neurais convolucionais, alto redshift , mapas de saliência, SC4K

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Lyman-α (Lyα) is a prominent spectral line in the ultraviolet region of the electromagnetic spectrum, corresponding to the transition of a hydrogen electron from its second energy level (n=2) to the ground state (n=1). This transition emits a photon with a wavelength of 1215.67 $\mathring{A}$ and represents one of the strongest emission lines in the Universe, appearing in the spectra of a diverse range of astrophysical objects [2, 3]. The significance of the Lyα line for extragalactic astrophysics and observational cosmology was recognised early, when Partridge and Peebles [4] predicted that primeval young galaxies would be strong Lyα emitters at very high redshifts (on the order of $z \approx$ 10–30). Although this redshift range was overestimated when compared to later observations($z \approx 11.4$) such as (e.g., Heintz et al. [5]), their insight was forward-looking: Lyα emission is a crucial tracer of young galaxies in the distant Universe (e.g., Nilsson [6], Huang et al. [7]).

Lyman-α emitters (LAEs) are galaxies, typically young star-forming systems or active galactic nuclei (AGN), that exhibit strong Lyα emission lines in their spectra [8]. In practice, LAEs are often identified by a rest-frame Lyα equivalent width above a certain threshold, commonly $EW_0 > 20$ Å [3, 8], which distinguishes them as objects dominated by this ultraviolet line. Most LAEs have been detected at high redshifts ($z \geq 2$). This is partly because the Lyα line, which has a rest-frame wavelength of 1215.67 $\mathring{A}$, is redshifted into the optical or near-infrared regime at $z > 2$, where ground-based and space-based instruments have optimal sensitivity. At lower redshifts ($z \leq 2$), however, the line remains in the far-ultraviolet, making it more difficult to observe from the ground due to atmospheric absorption. Furthermore, nearby galaxies, which dominate the low-redshift Universe, tend to contain significant amounts of dust and neutral hydrogen. These components strongly attenuate Lyα photons, further hindering their detection even when space-based facilities

are used. Consequently, strong Lyα emitters are comparatively rare in the nearby Universe (e.g., [3]) and are predominantly observed in the distant, early Universe (e.g., [9]).

Studying LAEs provides valuable insights into galaxy formation and evolution (see e.g., Partridge and Peebles [4]). Lyα emission is generally linked to energetic star formation, as hot, young stars ionise the surrounding hydrogen gas, which recombines and emits Lyα photons. LAEs therefore highlight sites of active star formation in young galaxies and can be utilized to trace how galaxies accumulate their stellar content over cosmic time (e.g., Oyarzún et al. [10]). By examining large samples of LAEs across different epochs, one can infer how the properties of star-forming galaxies (such as their star formation rates, masses, and gas content) evolve as the Universe ages (e.g., Mori and Umemura [11]). Moreover, Lyα line profiles can offer information about galactic kinematics (e.g., gas outflows) and the distribution of neutral gas in and around galaxies, as the interaction of Lyα photons with gas can broaden or shift the line (see e.g., [Hayes [3]).

LAEs also play a pivotal role in studying the intergalactic medium (IGM) (see e.g., Villasenor et al. [12], Nasir et al. [13]) and the Epoch of Reionisation (EoR) (see e.g., Dijkstra [14], Witten et al. [15]). The IGM refers to the diffuse gas found between galaxies, and during the EoR (approximately at redshifts $6 \leq z \leq 10$), this gas transitioned from being predominantly neutral to largely ionised (e.g., Gatuzz, E. et al. [16]). Lyα-emitting galaxies act as valuable probes of this cosmic transition. The presence or absence of Lyα emission from high-redshift galaxies is particularly sensitive to the ionization state of the surrounding IGM (e.g., Meiksin [17]). Consequently, studying LAEs across different redshifts, particularly as they approach the EoR($z \approx 6$), can provide crucial constraints on the timing and progression of reionization.

For decades, detecting Lyα-emitting galaxies proved challenging. It was only in the late 1990s that observational breakthroughs occurred. The first confirmed LAEs were reported after the development of deep imaging surveys and powerful telescopes. In 1996, Hu and McMahon [18] and Odewahn et al. [19] used narrowband imaging to identify high-z Lyα emitters, and shortly thereafter, the Hubble Deep Field and 8–10 meter-class ground telescopes enabled blank-field surveys for LAEs. By 1998, Cowie and Hu [20] and others had discovered many LAEs at redshifts $z > 2$. This pivotal advancement confirmed that galaxies with strong Lyα lines were indeed common in the young Universe and could be systematically found (e.g., [8]).

LAEs are predominantly observed at $z > 2$ due to the favourable redshifting of the

Lyα line into optical and near-infrared wavelengths. A Lyα photon emitted at 1216 Å will be observed at $\lambda_{obs} = 1216$ Å $(1 + z)$. For $z \approx 2$, this shifts the line to approximately 3648 Å (near the edge of the optical window), and for $z \approx 7$, it moves to $\approx$ 9700 Å (entering into the near-infrared). At these redshifts, Lyα falls within spectral ranges accessible to ground-based observatories equipped with sensitive CCDs and IR detectors (e.g., [21]). In contrast, Lyα from low-redshift galaxies (for example, $z \approx 0.3$ corresponds to $\approx$ 1600 Å) lies in the far-UV, requiring space-based instruments such as GALEX or HST (e.g., [22, 23]). Moreover, intrinsic galaxy properties at high redshift may favour strong Lyα emission: early galaxies tend to have low metallicities and less dust, reducing Lyα absorption (e.g., De Cia, A. et al. [24]) and often exhibit vigorous star formation (e.g., Su [25]). Nearby galaxies, particularly massive ones, typically contain more dust and fully ionised gas (e.g., Blanton and Moustakas [26]), which can suppress Lyα visibility. Consequently, most LAE surveys have targeted the high-redshift Universe, where both observational access and the likelihood of strong Lyα emission are greatest.

Identifying LAEs in practice relies on recognizing the Lyα line through imaging or spectroscopy. Two main techniques have been developed to find LAEs in the sky (see e.g., [8]):

- **Narrowband (NB) imaging:** This technique employs specialised narrowband filters to detect Lyα emission photometrically. The goal is to capture an image at a specific wavelength band adjusted for Lyα at a target redshift and compare it with broadband images in adjacent wavelengths of the same field. A galaxy exhibiting a strong Lyα line at the filter's wavelength will appear significantly brighter in the NB image than in a broadband image, which measures the continuum light (see e.g., [8]). Such an excess indicates the presence of an emission line, presumably Lyα. The central wavelength of the narrowband filter, λ_c, determines which redshift of Lyα is observed, following approximately $z \approx (\lambda_c/1216) - 1$. By selecting different NB filters, astronomers can target LAEs at specific redshifts. An important practical consideration is the atmospheric OH window - the Earth's atmosphere emits bright infrared sky lines (from OH molecules) that can overwhelm faint signals (see e.g., [27]). Therefore, NB filters are frequently placed in wavelength intervals between these sky lines (known as OH transparency windows) to minimise background noise (see e.g., [27]). Through narrowband imaging, large volumes can be surveyed relatively swiftly. For instance, Ouchi et al. [8] utilised a

NB921 filter (centred at $\approx 9200\mathring{A}$) to identify galaxies at $z \approx 6.6$, where a distinct NB excess uncovered candidate LAEs (e.g., cf. Figure 1.1). Narrowband surveys have proven very successful, providing samples of LAEs at discrete redshifts such as 2.2, 3.1, 4.6, 5.7, etc., which are often followed up with spectroscopy for confirmation (e.g., Sobral et al. [1]).

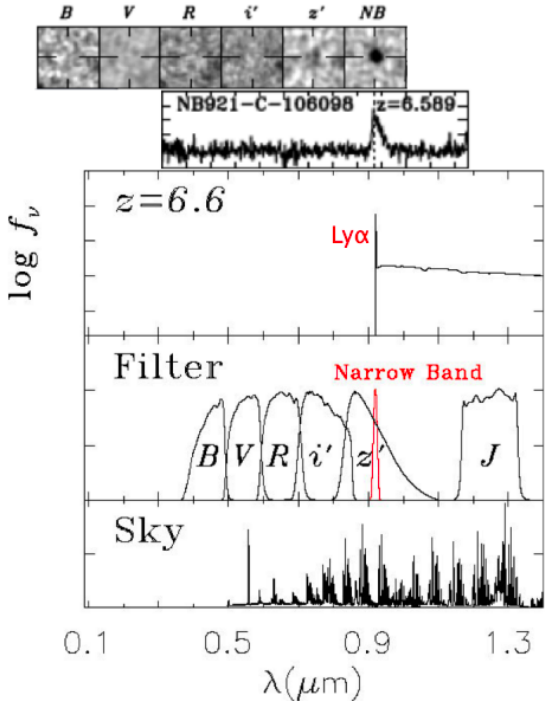


FIGURE 1.1: Image from Ouchi [28], narrowband selection for a LAE at $z = 6.6$. The top panel shows images of the LAE observed with broadbands (B, V, R, i', z') and a narrowband (NB921) filter at $\lambda_c \approx 9200$ Å. The second panel presents the spectrum of the LAE over the range $9050 - 9275$ Å. The third panel shows the model spectrum of an LAE redshifted to $z \approx 6.6$. The fourth panel shows transmission curves of the filters, and the bottom panel shows atmospheric OH lines.

- **Spectroscopic Searches:** LAEs can also be discovered via their spectral signatures without relying on narrowband pre-selection (e.g., [29, 30]). There are two approaches here: slitless and slit-based spectroscopy. Slitless spectroscopy involves taking a dispersed image of the sky (using a prism or grating, often called a grism) without any slit, so that every source in the field has its light spread into a spectrum. Instruments on space telescopes (like HST, JWST, or Euclid) and some ground-based setups can do this. LAEs then appear as objects showing an isolated emission line in the spectrum with no corresponding continuum, since many high-z LAEs are very faint in the continuum) (see e.g., [23, 31, 32]). The advantage is that one can blindly

search for emission lines across a wide field. However, on the ground this is challenging due to overlapping spectra and bright sky background; in space, where the sky background is much darker, slitless methods have successfully identified LAEs in deep fields (e.g., [33]). A more modern approach uses integral field spectrographs (IFS), such as the MUSE instrument on the VLT, which essentially take a spectrum at every position in a small field of view. IFS surveys allow a "blind" spectroscopic search for LAEs across the field without predefined targets (e.g., [34]). Each technique, narrowband imaging, grism, and IFS spectroscopy, has its own strengths, and together they have built a complementary picture of the LAE population across cosmic time (e.g., [8]).

Figure 1.2 displays an example of this technique application.

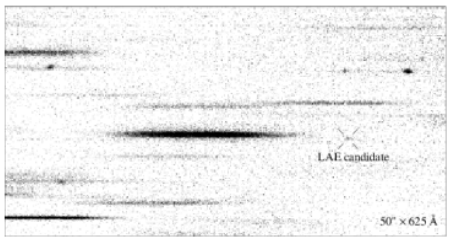


FIGURE 1.2: Image from Ouchi [28], showing an LAE candidate detected in VLT/FORS2 [35] grism data from a blank field.

Producing a strong Lyα line requires a source of ionizing photons and the right conditions for re-emission (e.g., [36]). Several physical processes can generate Lyα emission in galaxies and their environments. According to Ouchi et al. [8], the major sources of Lyα photons include:

1. **Young massive stars:** Hot OB-type stars within star-forming galaxies emit abundant ultraviolet radiation that excites or ionizes the surrounding interstellar medium. When the hydrogen lowers its energy or recombines, it can produce Lyα photons. LAEs often experience intense bursts of star formation that lead to this mechanism. [14, 37]

2. **Active Galactic Nuclei (AGN):** An accreting supermassive black hole (quasar or Seyfert nuclei) can photoionize gas and produce strong emission lines, including Lyα. Some LAEs may in fact be powered by AGN activity rather than purely by stars.

3. **Shock heating:** Outflows or supernova-driven winds can create shock fronts in the gas. Collisional excitation of neutral hydrogen in shocks (for example, in galactic superwinds) can lead to Lyα emission as the gas cools, a process often termed "shock heating" (see e.g., [38]).

4. **Infalling gas (cold accretion):** In the early Universe, gas falling into dark matter halos can form filamentary streams that reach the galaxy. As this infalling gas is heated and ionized at the interface with the galaxy, it may emit Lyα (often called "cold accretion" radiation; see e.g., [39]).

5. **Fluorescent illumination:** The gas in a galaxy's circumgalactic medium (CGM) or even the intergalactic medium can shine in Lyα if illuminated by an external source of ultraviolet photons. For example, the UV background radiation from quasars or star-forming galaxies can photoionize the neutral hydrogen in the CGM/IGM, causing it to fluoresce in Lyα (see e.g., [40]).

The first two sources (young stars and AGN) are internal to galaxies and pertain to the galaxy's interstellar medium. The latter three (shocks, accretion, and fluorescence) involve gas in the circumgalactic or intergalactic environment around galaxies. Multiple mechanisms might contribute simultaneously (see e.g., [8]). It is also worth noting that what we perceive as a single LAE could blend a central galaxy and smaller satellite galaxies or gas clumps; unresolved faint companions can add to the observed Lyα luminosity (see e.g., [8]). Identifying the dominant Lyα production mechanism for a given LAE can be challenging, but doing so is key to understanding the nature of these galaxies.

Although Lyα photons can be generated by multiple processes within galaxies, not all of them successfully escape the interstellar and itergalactic media to be observed by telescopes (e.g., [14]). Detecting Lyα emission is challenging primarily due to the resonant interactions between Lyα photons and neutral hydrogen [41]. Within galaxies, particularly young, gas-rich systems, the interstellar medium typically contains significant amounts of neutral hydrogen [42]. Lyα photons emitted near galaxy centers scatter repeatedly off these hydrogen atoms, significantly increasing their path length (e.g., [43]) This prolonged journey raises the likelihood that photons will be absorbed by dust, if present, or redirected far from their original emission points, altering their frequency. Consequently, galaxies actively forming stars (and thus producing Lyα photons) might appear faint in

Lyα if these photons become trapped by neutral gas and dust [44]. This problem intensifies at higher redshifts, where galaxies are more gas-rich, and near the EoR, where the IGM itself contains patches of neutral hydrogen [45]. Indeed, theoretical models have long indicated that Lyα photons generally cannot escape their host galaxies or the neutral early Universe without special conditions [46].

Observations, however, have shown that Lyα emission can escape, implying that galaxies have ways to let Lyα photons out [47]. Possible mechanisms for facilitating Lyα escape include galaxy-scale outflows that push neutral gas away from the center, creating channels of lower column density through which Lyα can leak out [48]. Outflows can also Doppler shift Lyα photons out of resonance with surrounding neutral gas, allowing them to traverse the ISM with less scattering [42]. Clumpy ISM structures (as opposed to a uniform distribution of gas and dust) might also permit higher escape fractions, since photons can scatter off clumps and dodge around denser regions [49]. There is ongoing research into the exact interplay of geometry, kinematics, and dust that yields a high Lyα transmission [50].

As noted, in the context of the early Universe, a LAE likely signals that its host galaxy is located in a partially ionized zone of the IGM. Radiation from the galaxy (or nearby galaxies) may have ionized a bubble in the IGM, which dramatically reduces the IGM opacity for Lyα. Still, even with such an ionized bubble, the galaxy must allow Lyα photons to escape from its immediate vicinity [51–53]. Why some galaxies are LAEs and others are not (even at similar epochs) remains an open question. It appears to depend on a complex combination of the galaxy's age, metallicity, dust content, gas kinematics, and environment. Previous studies have found correlations between Lyα luminosity or equivalent width and properties like star formation rate [54], UV luminosity [55], or halo mass [56], but these correlations exhibit scatter and exceptions [57].

Fundamentally, the escape of Lyα photons is a balance between production and destruction: the abundant production in young galaxies versus the multiple absorption and scattering processes that hinder the photons' escape. Overcoming these observational challenges often requires large statistical samples of LAEs (to average over cosmic variance and capture a range of conditions) and detailed follow-up of individual objects to piece together their story. Addressing these challenges is important not just for understanding Lyα emitters themselves, but also for what LAEs can tell us about cosmic history, such as reionization and galaxy evolution. Consequently, advancing our understanding

relies heavily on large-scale surveys and data-intensive astronomy to systematically study LAEs across various environments and epochs.

Astronomy has entered an era of "big data", marked by massive surveys and missions that gather unprecedented volumes of information. Over the past two decades, numerous large-scale projects have mapped the sky across various wavelengths, generating extensive catalogs of galaxies, stars, and other celestial objects. Space-based observatories like Euclid and the James Webb Space Telescope (JWST) exemplify this trend: Euclid, launched in 2023, is expected to survey billions of galaxies, producing nearly 1 petabyte of data per year [31], while JWST instruments generate hundreds of gigabits of data daily [32]. Ground-based efforts such as SDSS [58], Pan-STARRS [59], the Dark Energy Survey (DESI) [60], and the upcoming LSST at the Vera C. Rubin Observatory [61] further expand this data landscape. LSST alone will produce approximately 15 terabytes of data per night. Radio astronomy is no exception, with projects like the Square Kilometre Array (SKA) [62] expected to generate exabytes of data once fully operational. These numbers vastly exceed those from previous generations of surveys.

In this context, the search for LAEs is a double-edged sword. On one hand, large surveys are ideal for finding and studying significant numbers of LAEs. Surveys like COSMOS, with its 2 square degree multiwavelength coverage [63], have cataloged thousands of galaxy candidates [64], including many LAEs, providing rich datasets to analyze statistically. Faint LAEs, previously beyond detection thresholds, are now identifiable [65–67]. On the other hand, the sheer volume of the data makes manual identification strategies increasingly impractical. For example, narrowband imaging campaigns for LAEs require scanning large sky areas with specialized filters, producing millions of source detections that must be filtered for the telltale Lyα signature [68]. Spectroscopic surveys like zCOSMOS [69], DEIMOS [70], VUDS [71], and MUSE [72] yield catalogs of spectral lines that need identification. Traditionally, identifying LAEs in big datasets involved significant human effort visually inspecting color-color plots [73], images [74], or spectra [47] to pick out candidates, and straightforward cuts on data (e.g., selecting objects with a certain color excess in narrowband; [75]). This approach becomes impractical as datasets grow to billions of objects.

To meet these challenges, machine learning, especially deep learning, has emerged as a transformative tool. Convolutional Neural Networks (CNNs), in particular, are well-suited to classification tasks based on imaging data. Recent studies (e.g., Yoshioka et al.

76) have demonstrated that deep learning models can achieve true positive rates of 77% while maintaining low false positive rates of 14% in identifying LAEs, and can generalize well to new observations, including data from JWST. This kind of success paves the way for assembling large LAEs samples directly from broad photometric surveys, which is much more efficient than classic narrowband searches. Interestingly, beyond classification, such models can be used to probe astrophysical conditions: for instance, differences between the ML-predicted LAEs population and the spectroscopically confirmed LAEs can provide clues about the neutral gas distribution in the EoR. [76] showed that comparing their model's predictions with actual observations allowed them to place constraints on the typical sizes of ionized bubbles in the high-z IGM. This illustrates that big-data-driven machine learning models not only accelerate discovery, but can also yield insights into the underlying physics.

In light of the above developments, this dissertation leverages deep learning methods to improve the identification and analysis of Lyman-α emitters. The core goal is to develop a convolutional neural network model capable of recognizing LAEs from imaging data alone, and to use this model to investigate the conditions that enable Lyα escape. We specifically focus on using broadband photometric data (e.g., multi-band telescope images that can be combined into color images) as the input, so that our approach does not rely on narrowband filters or prior spectral information. The objectives of this work can be summarized as follows:

- **LAE Classification:** Train a CNN-based classifier to distinguish LAEs galaxies from non-LAEs using composite color images (constructed from broadband filters). The model will be designed to learn the subtle features associated with Lyα emission (such as color excesses or dropouts) and will be tested to ensure it generalizes well beyond the training sample, meaning it can reliably pick out LAEs in new, unseen datasets. A high classification accuracy is desired so that the resulting candidate lists are trustworthy for further study.

- **Physical Property Prediction:** Implement regression models to directly predict key physical properties of the galaxies identified as LAEs. In particular, I aim to estimate each galaxy's Lyα line luminosity, rest-frame equivalent width, and redshift from the same input imaging data. Successfully predicting these properties would demonstrate that the network is capturing not just a yes/no classification, but also information related to the strength of Lyα emission line and the galaxy's distance.

- **Interpretability and Physical Insights:** Beyond performance, I will probe the interpretability of the trained neural network to understand what features in the data drive its decisions. Techniques such as visualization of neuron activation (e.g., saliency maps) will be used to see which parts of an image are most influential in identifying an LAE. By doing so, I hope to connect the network's internal logic with physical aspects like the presence of a Lyman-break, and the spatial extent of UV light, which might relate their compactness that allows Lyα escape.

To achieve these goals, I structure the dissertation as follows. **Chapter 1 (Introduction)** has provided the background and motivation, outlining the importance of LAEs, the observational methods and challenges in detecting them, and the emergence of big data and deep learning techniques to study them. **Chapter 2 (Data)** describes the datasets used in this work, including the imaging surveys and LAE catalogs from which my training and testing samples are drawn. I detail the characteristics of these data (such as filters, depths, and redshift coverage) and the preprocessing steps taken to make them suitable for input into the neural network (for example, image cutout preparation, photometric corrections, and augmentation). **Chapter 3 (Methods)** explains the deep learning models and techniques employed. I discuss the CNN architecture designed for classification, the loss functions and optimization strategy, and how I set up the regression models for properties prediction. This chapter also covers the evaluation metrics used to assess performance (accuracy, precision, recall for classification and mean-absolute-error, mean square Error for regression) and describes any cross-validation or hyperparameter tuning performed. **Chapter 4 (Results)** presents the outcomes of my experiments. I report the classifier's performance in identifying LAEs and the accuracy of the property predictions, and I compare the results with sources that have spectroscopic confirmation. To interpret the model's decision-making, I employ visualization tools such as saliency maps, which highlight the regions of input images most influential to the CNN. I also examine misclassified examples to identify and understand the model's limitations. **Chapter 5 (Discussion)** discusses the interpretation of findings, model limitations, and their scientific implications. Finally, **Chapter 6: (Conclusion)** summarizes the key findings of the dissertation, reflecting on how deep learning can advanced the study of LAEs. I highlight the contributions of this work to the broader field such as providing a new tool for LAE selection and insights into Lyα escape. All tables and the full source code used in this study are openly available at https://github.com/Onirb/Tese, ensuring reproducibility and enabling future

research.

Throughout this work, I assume a flat ΛCDM cosmology with $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, $\Omega_M = 0.3$, and $\Omega_\Lambda = 0.7$ [77].

Chapter 2

Data

The data used in this study originates from three primary sources: two for tabular data and one for imaging. The SC4K survey [1] and the COSMOS 2020 catalog [64] were selected due to the exceptional data quality, extensive multi-wavelength coverage, and reliable photometric redshift estimates provided by the COSMOS field, which is particularly valuable as it hosts the largest, most homogeneous, and consistently defined sample of LAEs available to date making it the optimal environment to study LAEs using machine learning techniques.

To ensure model robustness and mitigate class imbalance, we constructed balanced datasets with respect to redshift, i-band magnitude distribution (within the bin sizes defined in Section 2.2), and total number of sources. This resulted in one dataset of LAEs, selected from the SC4K [1] survey in the COSMOS field, and seven comparison datasets composed of other star-forming galaxies (oSFG) from COSMOS2020 [64], which we refer to as non-LAEs (nLAEs). These nLAEs were not identified as LAEs by SC4K, and are matched to the LAE sample in redshift and i-band magnitude distributions. While some of these sources may contain undetected or low-EW Lyα emission, we treat them statistically as typical star-forming galaxies for the purpose of training and validation. Each dataset contains 3317 sources. A separate prediction dataset was also created using the remaining COSMOS2020 sources which are unlabeled. The details of this procedure are described in Section 2.2.

For imaging data, RGB images were generated using observations from the Subaru Suprime-Cam [78], an imaging camera installed on the 8.2m Subaru telescope. The Suprime-Cam offers a wide field of view of 1.5 degrees in diameter and is equipped with a variety of filters, including 5 broadband filters (g, r, i, z, y), 4 narrow-band filters, and several

medium-band filters. For this work, we specifically used the homogenized broadband filters g+, r+, and i+ from the HSC Subaru to construct RGB images, the choice of these 3 is first 3 filters are need for a RGB and these GRI cover the usual narrowband filters used to study LAEs.

2.1 Catalogs and surveys

2.1.1 COSMOS2020

The COSMOS2020 catalog includes two major photometric versions: Classic and Farmer. While the Farmer catalog provides profile-fitting photometry with higher precision in localized regions, its coverage is limited by mask restrictions and constrained primarily to the UltraVISTA footprint. For this study, the Classic version is adopted, which offers reliable aperture photometry and ensures complete and uniform spatial coverage across the full COSMOS field. This choice facilitates consistent comparison with the SC4K sample, which also spans the full field [64].

The Cosmic Evolution Survey (COSMOS) 2020 [64] is pivotal in extragalactic astronomy, offering a comprehensive view of cosmic evolution. It details the meticulous collection, processing, and analysis of imaging data within the COSMOS field, spanning approximately 2 square degrees. This effort resulted in the creation of a refined reference photometric Redshift catalog, encompassing a vast ensemble of 1.7 million sources.

The catalog encompasses a rich array of multi-wavelength photometry data, with nearly 966,000 sources measured across all available broadband data. The photometry extraction process involved the utilization of traditional aperture photometric methods alongside a profile-fitting photometric extraction tool termed ”The Farmer”, developed specifically for that study.

Figure 2.1 shows the transmission curves of the filters from the COSMOS2020 dataset. The three broadband filters used to construct the RGB images are highlighted in bright colors, with their curves normalized to a maximum transmission of 1.0. The remaining broadband filters are also shown in color but with lower opacity to indicate they were not used for image construction. In contrast, all medium- and narrow-band filters are displayed in grey and their transmission curves are normalized to a maximum of 0.3.

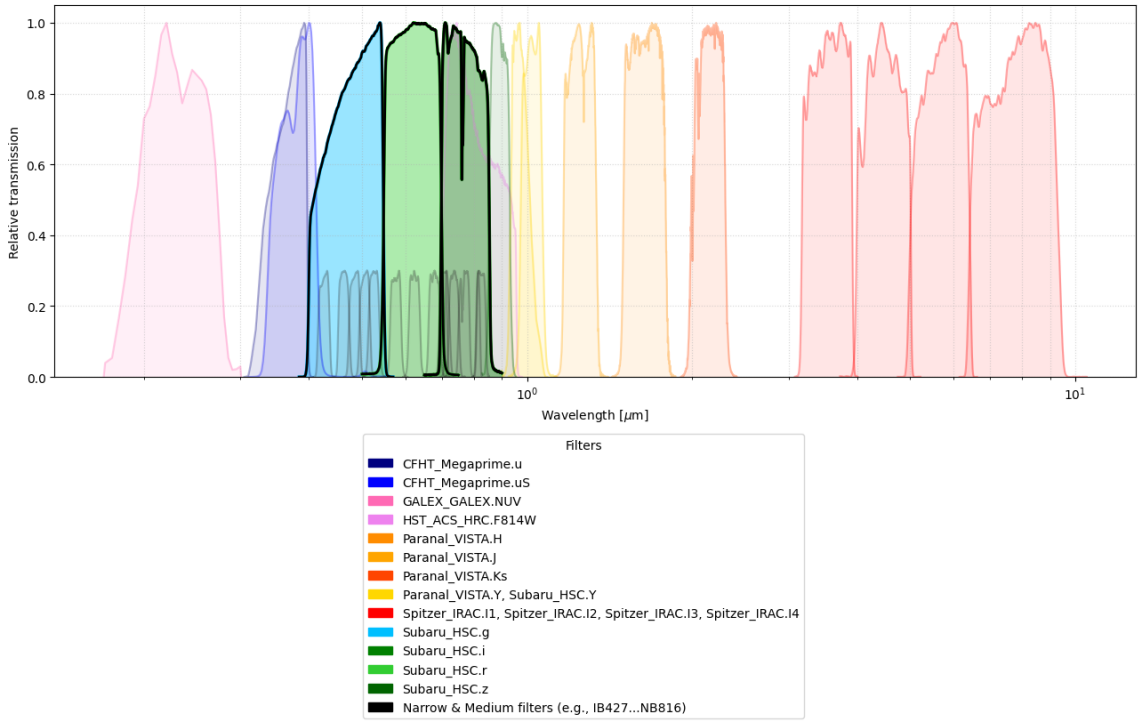


FIGURE 2.1: Normalized transmission curves of filters used in the COSMOS2020 and SC4K surveys. Broadband filters are displayed in color and normalized to 1, while narrow- and medium-band filters are shown in gray and normalized to 0.3. The three broadband filters used to construct the RGB images (g+, r+, i+) are highlighted with a thicker black contour and increased opacity.

2.1.2 SC4K

Description of the SC4K: The Slicing the Cosmos 4k (SC4k) [1] survey, selects a total of 3908 LAEs using narrow and medium bands techniques over the COSMOS field, their selection is inside a redshift range of 2 to 6, passing by the star formation peak, until the re-ionization era.

They apply two criteria: one to select the Lyman break and the other to remove stars or red galaxies that have a strong Balmer break that mimics the Lyman break, in agreement with [79].

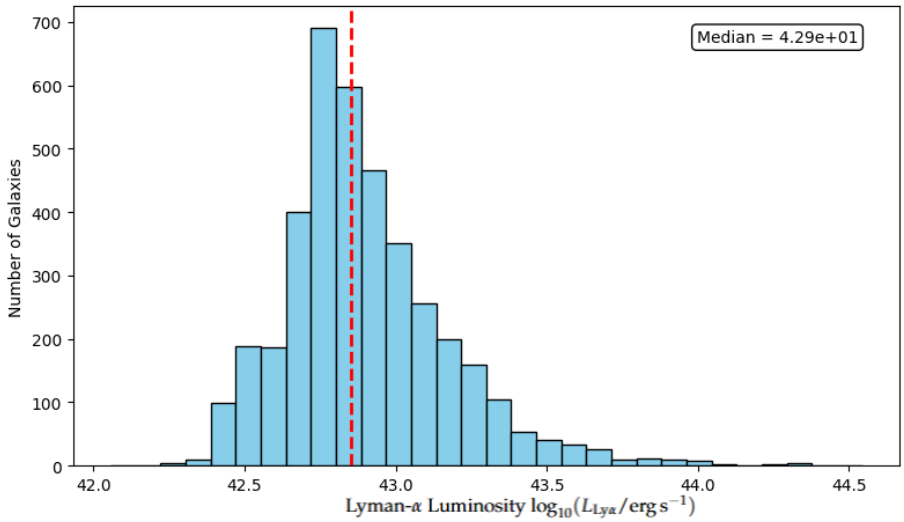
The SC4k survey has a total of 12 medium bands and 4 narrow bands, each band refers to a redshift slice, and we use this value to track the redshift of each source. Table 2.1 shows the number of LAE in each redshift slice, for a total of 3908 LAE, it also displays the amount within each redshift slice for the LAE Dataset generated from this survey, a total of 3317 (see Section 2.2).

The Figure 2.2 displays the distribution for the log_{10} of Lyman Alpha luminosity and the log_{10} of Equivalent width for the SC4K survey.

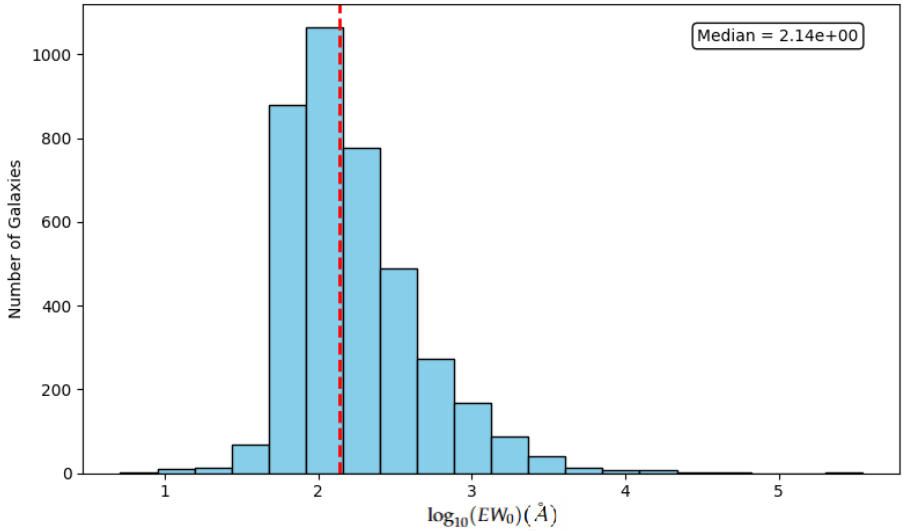
TABLE 2.1: Adapted from [1], Describe the filter, redshift range, and the number of LAE candidates for each slice, a total of 12 medium bands and 4 narrow bands. The column LAE dataset refers to the values obtained for this dissertation after the data processing, while the others columns are the original values used in [1]

Selection Filter	**Lyα Redshift range**	**# LAE Candidates**	**# LAE Dataset**
IA427	2.42 – 2.59	741	679
IA464	2.72 – 2.90	311	294
IA484	2.92 – 3.10	483	621
IA505	3.07 – 3.26	478	416
IA527	3.30 – 3.50	641	578
IA574	3.63 – 3.85	98	92
IA624	4.00 – 4.16	124	124
IA679	4.39 – 4.57	79	71
IA709	4.53 – 4.72	176	74
IA738	4.67 – 4.87	90	62
IA767	4.81 – 5.01	99	31
IA827	5.64 – 5.92	55	20
NB392	2.20 – 2.24	159	85
NB501	3.08 – 3.16	159	13
NB711	4.83 – 4.89	78	48
NB816	5.65 – 5.75	192	109
TOTAL		**3908**	**3317**

The SC4K survey provides a curated selection of LAEs using narrow and medium-band filters, but it only covers specific redshift slices and relies on visual inspection for source validation. In contrast, the COSMOS2020 catalog offers continuous multi-band photometry over the entire COSMOS field. This difference in methodology and coverage makes COSMOS2020 an essential complementary dataset for both enriching the LAE feature set and constructing a robust comparison sample of nLAE.



(A) Histogram of log_{10} of Lyman alpha Luminosity from the SC4K data



(B) Histogram of $\log_{10}(EW_0)$ from the SC4K data

FIGURE 2.2: Distributions of Equivalent Width ($\log_{10}(EW_0)$) and Lyman Alpha Luminosity ($L_{Ly\alpha}$) of the SC4K data.

2.2 Data processing

The LAE dataset was constructed using the SC4K survey [1] for the initial selection of LAEs, cross-matched with the COSMOS2020 catalog [64] to obtain additional photometric features. The cross-matching was performed using TOPCAT [80], based on Right Ascension (RA) and Declination (Dec), using the Sky algorithm with a maximum error of 1 arcsecond. The matching strategy was symmetric, selecting the best match. This process resulted in 3346 matched sources.

To reduce contamination from AGNs, we applied a cut on the COSMOS2020 dataset column ip_mag_APER3, keeping only sources with $ip\_mag\_APER3 > 22$, following the analysis from Calhau et al. [81], which states that from the SC4K analisys most of the AGN are located bellow these values. After this filtering, the final LAE dataset consisted of 3317 sources. These objects retained all SC4K features, now complemented by COSMOS2020 measurements. The dataset was then divided into training, testing, and validation subsets. The last column in the Table 2.1 displays the number for each readshift bin.

The comparison datasets of nLAEs were created from the COSMOS 2020 catalog, specifically, the Classic Catalog, where the photometry aperture is performed on PSF-homogenized images [82], [83]. We excluded all 3317 LAE sources from COSMOS2020 and applied the following selection criteria: photometric redshift in the range $2 < z < 6$ and $0 < ip\_mag\_APER3 < 40$. These limits prevent the inclusion of sources with invalid magnitudes (e.g., -1 or 99.9, used to encode missing values in astronomical catalogs).

To ensure the nLAE datasets had the same redshift and magnitude distributions as the LAE sample, we defined bins in redshift (z) from 2 to 6 with step 1, and in ip_mag_APER3 from 22 to the LAE dataset maximum (≈ 40), with step 0.5. For each bin combination, we counted the number of LAEs and randomly selected the same number of nLAE sources. This sampling strategy was repeated seven times, generating seven independent nLAE datasets, each with 3317 galaxies and identical distribution profiles to the LAE dataset.

The remaining COSMOS2020 sources, after removing the LAE dataset and the first nLAE dataset, were grouped into a prediction dataset containing 191,826 unlabeled sources.

Figure 2.3 illustrates the redshift and ip_mag_APER3 distribution for the LAE and the seven nLAE datasets. The LAE sample is shown in green with bars and dots, and each nLAE dataset is represented by a unique color and marker style, with hatched histograms indicating the shared distribution shape.

Although three types of datasets are available, LAE, nLAE, and prediction, only classification analysis uses them, while regression analysis is conducted exclusively on the LAE dataset, which provides well-constrained target features: redshift, Lyman Alpha Luminosity, and Equivalent Width (EW_0). To optimize the performance of the Convolutional Neural Network, appropriate preprocessing is applied: Lyman Alpha Luminosity is scaled using a RobustScaler [84], and the Equivalent Width is log-transformed using base 10.

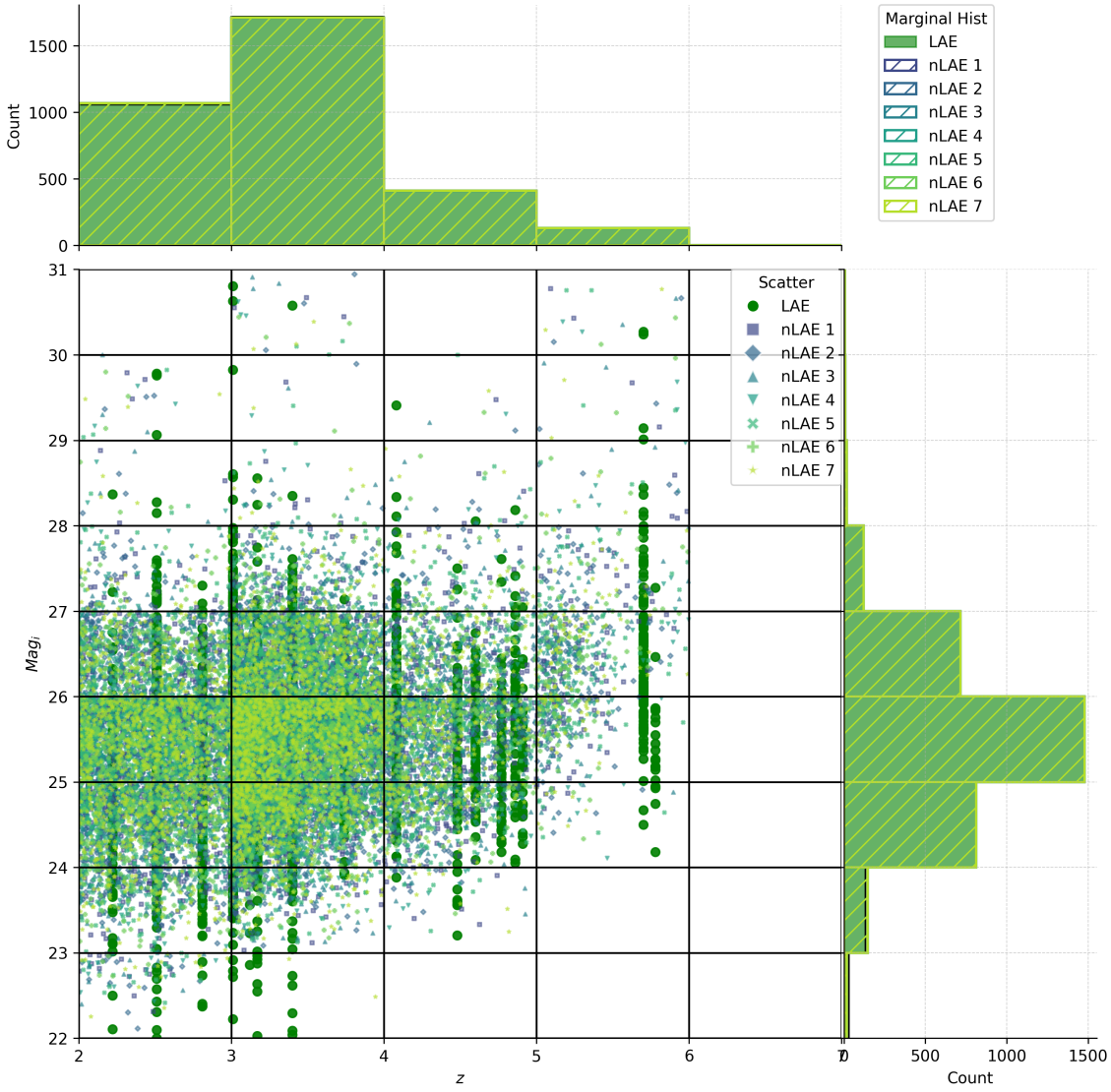


FIGURE 2.3: The plot shows the distribution of I-band magnitude as a function of redshift for all samples, including 1 LAE and 7 nLAEs.

Additionally, three histograms are presented in Figures 2.4 - 2.6 to provide a detailed overview of the distributions of the key physical features used in the regression analysis. These values are provinient from the LAE dataset values. Figure 2.4 shows the redshift distribution of LAEs, while Figures 2.5 and 2.6 display the distributions of the log-transformed equivalent width ($\log_{10}(EW_0)$) and the scaled Lyman-alpha luminosity ($log_{10}(L_{Ly\alpha})$), respectively. These plots confirm that the LAE dataset spans a wide range of physical properties, with statistically meaningful variation in all parameters.

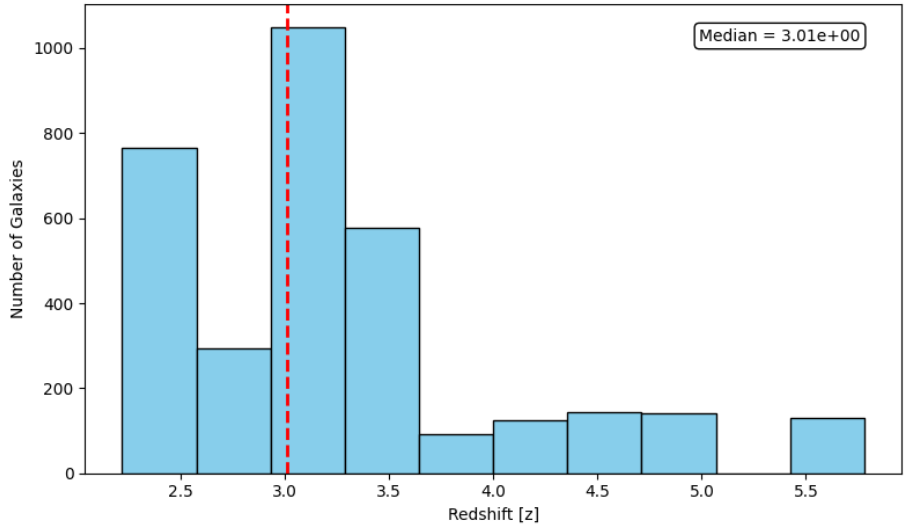


FIGURE 2.4: Redshift Distribution: The histogram shows the distribution of redshift values in the LAE dataset, with prominent peaks around 2.5 and 3.0, and smaller peaks beyond 3.5.

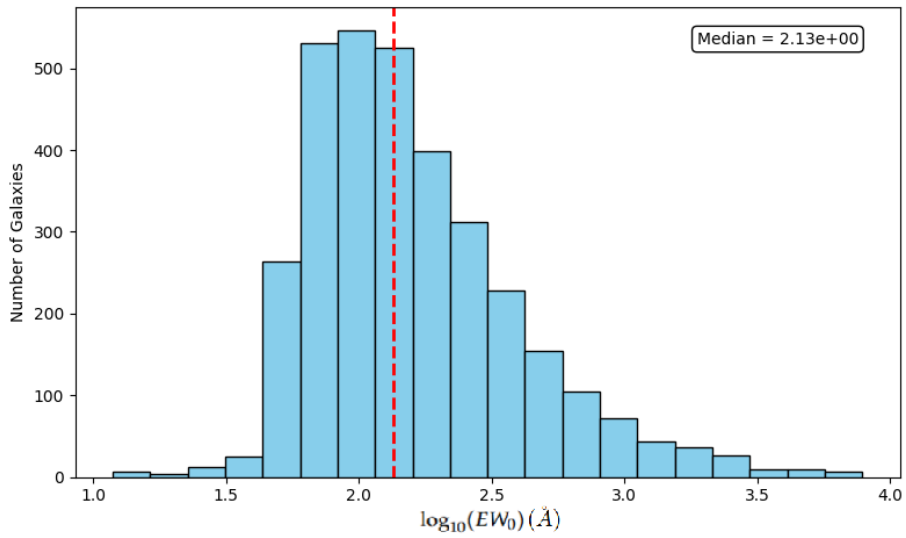


FIGURE 2.5: LogEW0 histogram: Displays the distribution of log-transformed Equivalent Width ($\log_{10}(EW_0)$) of the Lyα line values from the the LAE dataset, which is skewed to the left.

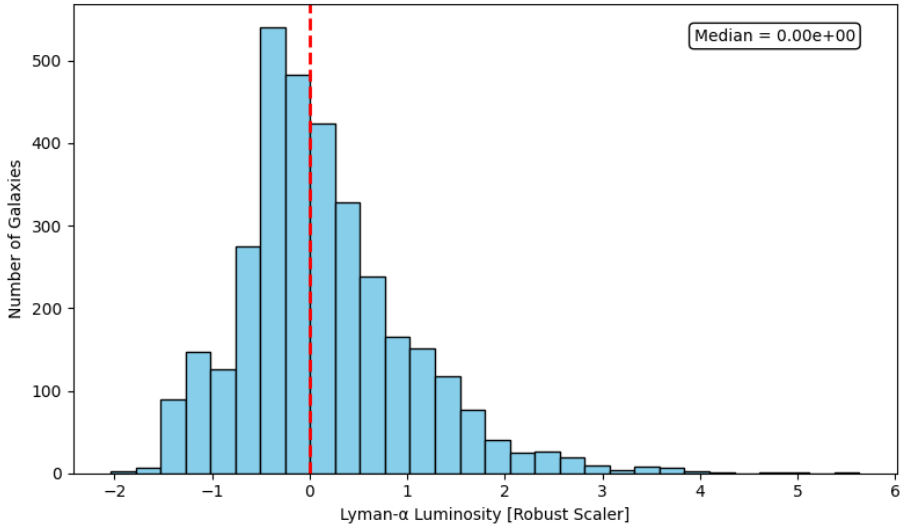
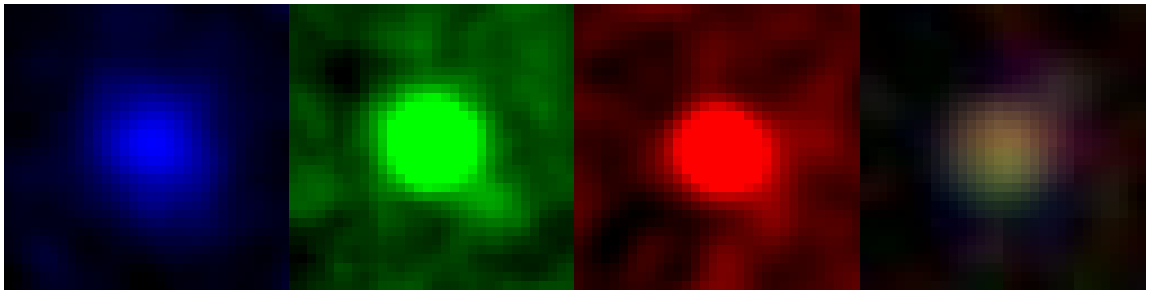


FIGURE 2.6: $L_{Ly\alpha}$ distribution: Illustrates the distribution of scaled Lyman Alpha Luminosity (LyaLum_scaled) values of the LAE dataset after scaling, with a peak around the mean and a long tail towards higher values.

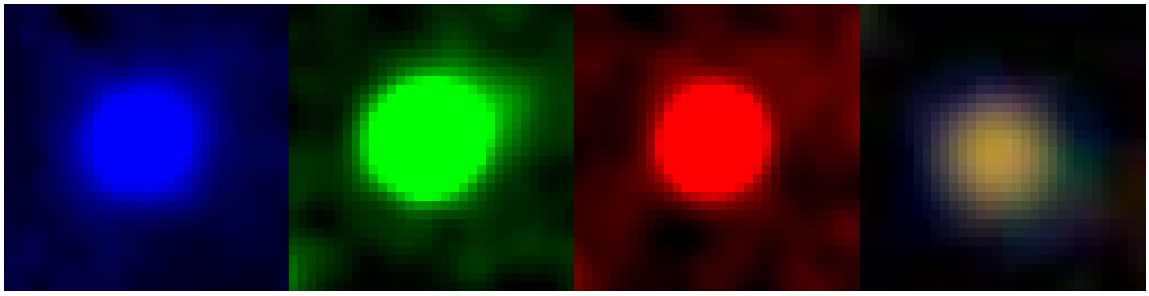
2.3 Image construction

We choose to work with the broadband G, R, and I filters, using the homogenized PSF mosaics from [78]. These broadband filters were selected because they provide high signal-to-noise ratios in the Subaru images and cover the rest-frame UV and optical regions most relevant to LAE morphology at redshifts 2 to 6. Figure 2.1 illustrates the transmission ranges of these filters. To construct the RGB images, we used the `Astropy` package [85] to extract the corresponding FITS cutouts from the mosaic images. These cutouts were then combined using the `make_lupton` function [86] with default parameters to produce RGB JPEG images of size 32×32 pixels.

We chose a final image size of 32x32 pixels based on empirical testing that showed no performance improvement when using 64x64. The reduced size limits the inclusion of background noise while preserving the key morphological features of the sources. The RGB image construction uses g+ as blue, r+ as green, and i+ as red, following standard astronomical color mapping conventions. The resulting images are shown in Fig 2.7 together with the filters in which the first row is a LAE source with redshift 2.81 while the second contains nLAE source with redshift 2.41.



(A) GP (blue) (B) RP (green) (C) IP (red) (D) LAE: RGB



(E) GP (blue) (F) RP (green) (G) IP (red) (H) nLAE: RGB

FIGURE 2.7: Example of RGB image used as final input for the CNN models, the first-row are LAE sources, while the second are nLAE sources

Chapter 3

Methods

This chapter presents the methodological framework that supports the analysis carried out in this work. The chapter is divided into two main sections. The first, Section 3.1, introduces the core principles of deep learning relevant to this study, including the architecture of CNNs, the activation functions employed, and the evaluation metrics used for both classification and regression tasks. This theoretical foundation is essential to understanding the motivations behind the choices made during model design and training.

The second part, Section 3.2, outlines the complete experimental methodology, from the initial classification setup to the regression models developed for predicting key astrophysical properties. Different training strategies, such as fine-tuning, model comparison, and the use of explainability tools like saliency maps, are discussed in detail. Moreover, alternative regression approaches are presented, including single-model multitarget prediction, independent models per target, and a chained regression pipeline using redshift as auxiliary input.

Together, these sections provide a comprehensive view of the techniques and reasoning that guided the implementation and evaluation of the models used in this research. The methodological choices were shaped by the specific challenges of working with astronomical image data, such as the need for interpretability, generalization across redshift ranges, and accurate estimation of galaxy properties from limited pixel information.

3.1 Deep learning

Machine Learning (ML) is a subset of Artificial Intelligence (AI) that focuses on developing algorithms that enable computers to learn from and make predictions based on data [87]. Unlike traditional programming, where explicit instructions are coded, machine learning models are trained on data to recognize patterns and make decisions with minimal human intervention. This approach allows for automating complex tasks and extracting meaningful insights from large datasets.

AI is the broader field that encompasses Machine Learning. AI aims to create systems capable of performing tasks that typically require human intelligence, such as understanding natural language, recognizing images, and making decisions [88]. AI includes various subfields, including natural language processing, robotics, computer vision, and more. Machine Learning represents a significant advancement within AI, providing powerful tools for data analysis and prediction.

A particularly influential subset of Machine Learning is Deep Learning (DL), which uses neural networks with many layers, known as deep neural networks, to model complex patterns in data [89]. One of the most effective types of DL models for image and spatial data is the CNN. They work by applying convolutional filters to input data, capturing spatial hierarchies and features at different levels of abstraction [90]. They consist of multiple layers, including convolutional layers, pooling layers, and fully connected layers, each playing a crucial role in feature extraction and pattern recognition.

The relationship between these fields is illustrated in Figure 3.1, which shows how AI encompasses ML, and within ML, DL forms a specialized subset that focuses on deep neural networks.

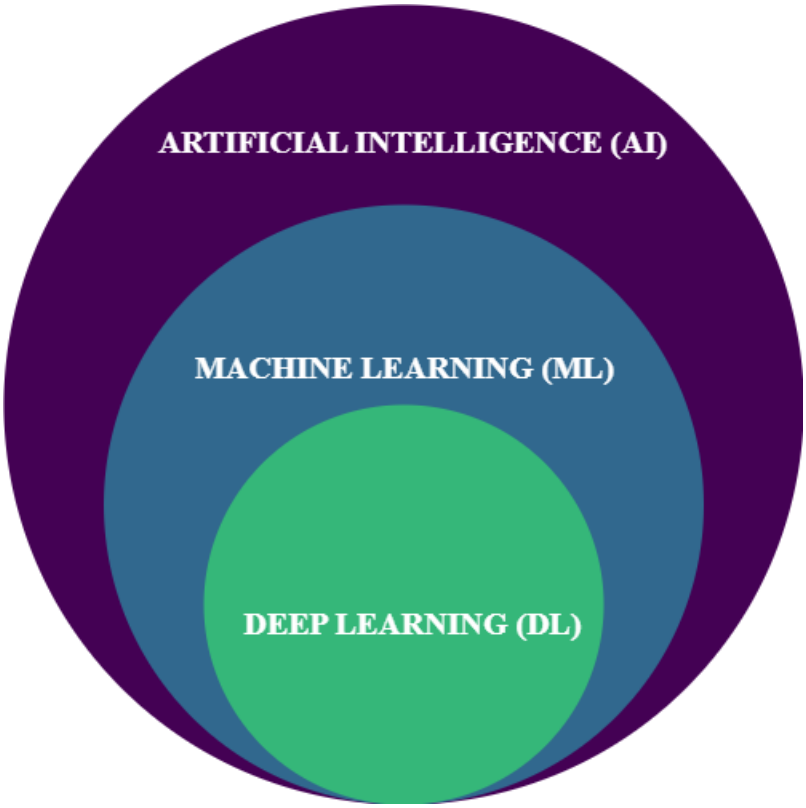


FIGURE 3.1: Hierarchical relationship between Artificial Intelligence (AI): Develops systems capable of performing tasks that normally require human intelligence, such as understanding natural language, recognising images and making decisions; Machine Learning (ML): Develops algorithms that enable computers to learn from data and make predictions, and Deep Learning (DL): Uses multilayer neural networks to model complex patterns in data. AI includes ML, which focuses on developing algorithms that enable computers to learn from data. Within ML, DL uses deep neural networks to model complex patterns. Figure from author.

3.1.1 Architecture

CNNs are a class of deep learning models designed to process data with grid-like topology, such as images [91]. They are composed of multiple layers, each serving a specific purpose and contributing to the overall ability of the network to learn and make predictions. The primary layers in a CNN include convolutional layers, activation functions, pooling layers, and fully connected layers, each playing a distinct role in the feature extraction and classification process.

Convolutional Layers are the foundational components of CNNs. These layers apply convolutional filters to the input image to detect local features such as edges, textures, and patterns. The convolution operation involves sliding a filter (or kernel) over the input data and computing dot products between the filter and local regions of the input [91]. This process generates feature maps, which highlight the presence of specific features within

the image. By stacking multiple convolutional layers, the network can detect increasingly complex and abstract features.

Following the convolutional layers, activation functions are applied to introduce non-linearity into the model. One of the most commonly used activation functions is the Rectified Linear Unit (ReLU). The ReLU function activates a node only if the input is above a certain threshold, effectively allowing the model to capture and learn more complex patterns in the data. Without these non-linear activation functions, the network would be limited to learning only linear relationships.

Pooling layers are used to downsample the feature maps, reducing their dimensionality and computational load while preserving important features. Pooling helps to make the model more robust to variations in the position of the features within the input image. Common pooling methods include max pooling, which selects the maximum value from a pooling window, and average pooling, which calculates the average value. These operations reduce the spatial dimensions of the feature maps, thus condensing the information and mitigating overfitting.

Towards the end of the CNN architecture, fully connected layers are employed. These layers function similarly to those in traditional neural networks and serve to combine the features extracted by the convolutional and pooling layers. The output of the final pooling or convolution layer is flattened and fed into one or more fully connected layers, which integrate the features and produce the final output predictions. These layers are crucial for the final decision-making process of the network.

Each layer in a CNN has parameters, such as weights and biases, which are learned during the training process. Training a CNN involves adjusting these parameters to minimize the difference between the predicted output and the true labels using a loss function. This optimization is typically achieved through a process called backpropagation [92, 93], where the network propagates the error gradient backward from the output layer to the input layer, updating the parameters to reduce the overall error.

Overall, the architecture of CNNs allows for efficient and effective feature extraction, making them particularly well-suited for tasks involving image recognition and other forms of spatial data analysis. By stacking multiple layers, each with a specific function, CNNs can learn and model complex patterns in the data, leading to highly accurate predictions.

3.1.2 Activation functions in neural networks

Activation functions are crucial components of neural network architecture, allowing the model to learn and represent complex patterns in data through the introduction of non-linearity. This subsection provides an overview of four widely used activation functions applied at various stages in this project: ReLU, sigmoid, softmax, and linear. Each function's definition, mechanism, and application within the layers of a CNN are detailed.

3.1.2.1 ReLU (rectified linear unit)

The ReLU activation function is defined as:

$$\text{ReLU}(x) = \max(0, x). \tag{3.1}$$

Popularized by Nair and Hinton [94], ReLU is favored in CNNs for its simplicity and efficiency. It addresses the vanishing gradient problem, allowing models to converge more quickly and achieve better performance in deep networks. By activating neurons only when the input is positive, ReLU introduces sparsity in the network, enhancing computational efficiency. Typically applied after convolutional layers, ReLU helps the model capture more intricate features by introducing non-linearity, promoting faster training speeds, and mitigating gradient-related issues.

3.1.2.2 Sigmoid

The sigmoid activation function is given by:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{3.2}$$

Historically significant, the sigmoid function was widely used in early neural networks [95]. It compresses input values to a range between 0 and 1, making it suitable for binary classification tasks. However, sigmoid functions can suffer from vanishing gradients, which can hinder learning in deeper networks. Often employed in the output layer of binary classifiers, the sigmoid function converts the network's output into a probability score, indicating class membership and facilitating the interpretation of binary classification predictions.

3.1.2.3 Softmax

The softmax activation function is defined as:

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}. \tag{3.3}$$

Widely used for multi-class classification problems [96], softmax transforms logits (raw prediction scores) into a probability distribution over multiple classes, ensuring that the sum of all probabilities equals one. Applied in the output layer of neural networks designed for multi-class classification, softmax allows the model to produce a probability distribution across different classes. This is crucial for tasks where each instance must be assigned to a single class among many, aiding in decision-making by providing interpretable probability scores.

3.1.2.4 Linear

The linear activation function is simply defined as:

$$f(x) = x. \tag{3.4}$$

Used in scenarios where the network's output should be a continuous value without non-linear transformation, the linear function, while not an activation function in the traditional sense, is essential for specific prediction types. Commonly utilized in the output layer of regression models, the linear activation function enables the model to produce a wide range of output values, making it suitable for tasks that require continuous and unbounded predictions.

Each layer in a CNN, including those utilizing these activation functions, has parameters (weights and biases) that are learned during the training process. Training involves adjusting these parameters to minimize the difference between the predicted output and the true labels using a loss function, often through a process called backpropagation.

These activation functions collectively enable CNNs to learn from complex datasets and perform a wide range of tasks, from binary and multi-class classification to continuous value predictions.

3.1.3 Metrics

Evaluating the performance of Machine Learning models requires the use of appropriate metrics that offer insights into the models' effectiveness. This section outlines several commonly used evaluation metrics applied at various stages in our project: accuracy, precision, recall, F1-Score, mean absolute error (MAE), root mean square error (RMSE), and mean squared error (MSE). Each metric is defined and discussed in terms of its strengths, weaknesses, and typical applications.

Accuracy is the most straightforward metric and is defined as the ratio of correctly predicted instances to the total number of instances:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}. \tag{3.5}$$

Accuracy provides a general measure of how often a classifier is correct. Its simplicity makes it popular for balanced datasets, where each class has roughly the same number of instances. However, accuracy can be misleading in imbalanced datasets. For instance, if one class is much more prevalent, a model that always predicts the majority class will achieve high accuracy despite failing to identify the minority class effectively.

Precision, or positive predictive value, is defined as the ratio of true positive predictions to the sum of true positive and false positive predictions:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}. \tag{3.6}$$

Precision measures the accuracy of positive predictions made by the model. It is particularly useful when the cost of false positives is high, such as in spam detection where incorrectly labeling legitimate emails as spam is undesirable. While precision focuses on the reliability of positive predictions, it does not account for false negatives, which can be problematic in scenarios where missing a positive instance is critical.

Recall, also known as sensitivity or the true positive rate, is defined as the ratio of true positive predictions to the sum of true positives and false negatives:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{3.7}$$

Recall measures the model's ability to identify all relevant instances within the dataset. It is crucial in scenarios where identifying all positive instances is essential, such as in

medical screenings or fraud detection. High recall often results in a decrease in precision, as the model may generate more false positives to capture all true positives.

F1-Score is the harmonic mean of precision and recall, providing a single metric that balances both:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3.8}$$

The F1-Score is particularly useful in cases of imbalanced class distribution where both precision and recall are important. It combines the strengths of both metrics into a single value, making it valuable when a balance between false positives and false negatives is desired. However, the F1-Score does not differentiate between the relative importance of precision and recall, which may be critical in some applications.

For the classification tasks discussed in this dissertation, accuracy was primarily used to evaluate the results, given the balanced nature of the dataset. However, other metrics were also computed, and the F1 Score was considered in specific cases.

Regarding the regression metrics:

Mean Absolute Error (MAE) is defined as the average of the absolute differences between predicted and actual values:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i|, \tag{3.9}$$

where x_i represents the predicted value and y_i denotes the actual value for the i-th instance, and n is the total number of instances. MAE is easy to interpret, reflecting the average magnitude of errors in the units of the original data. It treats all errors equally, which may not be ideal when larger errors should be penalized more [97]. Unlike squared error metrics, MAE is less sensitive to outliers.

Root Mean Square Error (RMSE) is the square root of the average of the squared differences between predicted and actual values:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2}. \tag{3.10}$$

RMSE provides a measure of the average magnitude of error, heavily penalizing larger deviations. Its sensitivity to large errors makes it suitable for scenarios where large deviations are particularly undesirable. RMSE also shares the same units as the original data,

facilitating interpretation [98]. However, its sensitivity to outliers can be a disadvantage when dealing with noisy datasets.

Mean Squared Error (MSE) is defined as the average of the squared differences between predicted and actual values:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2. \tag{3.11}$$

MSE measures the variance of prediction errors and penalizes larger errors more than smaller ones [99]. It is useful when large errors are particularly detrimental and is also mathematically convenient for optimization due to its differentiability. Like RMSE, MSE is sensitive to outliers and its squared units can make interpretation less intuitive compared to MAE.

3.2 Methodology overview

With the steps made in Section-2, the data is ready for classification and regression tasks, Figure 3.2 is a flow chart to clarify the structure of the data until this step.

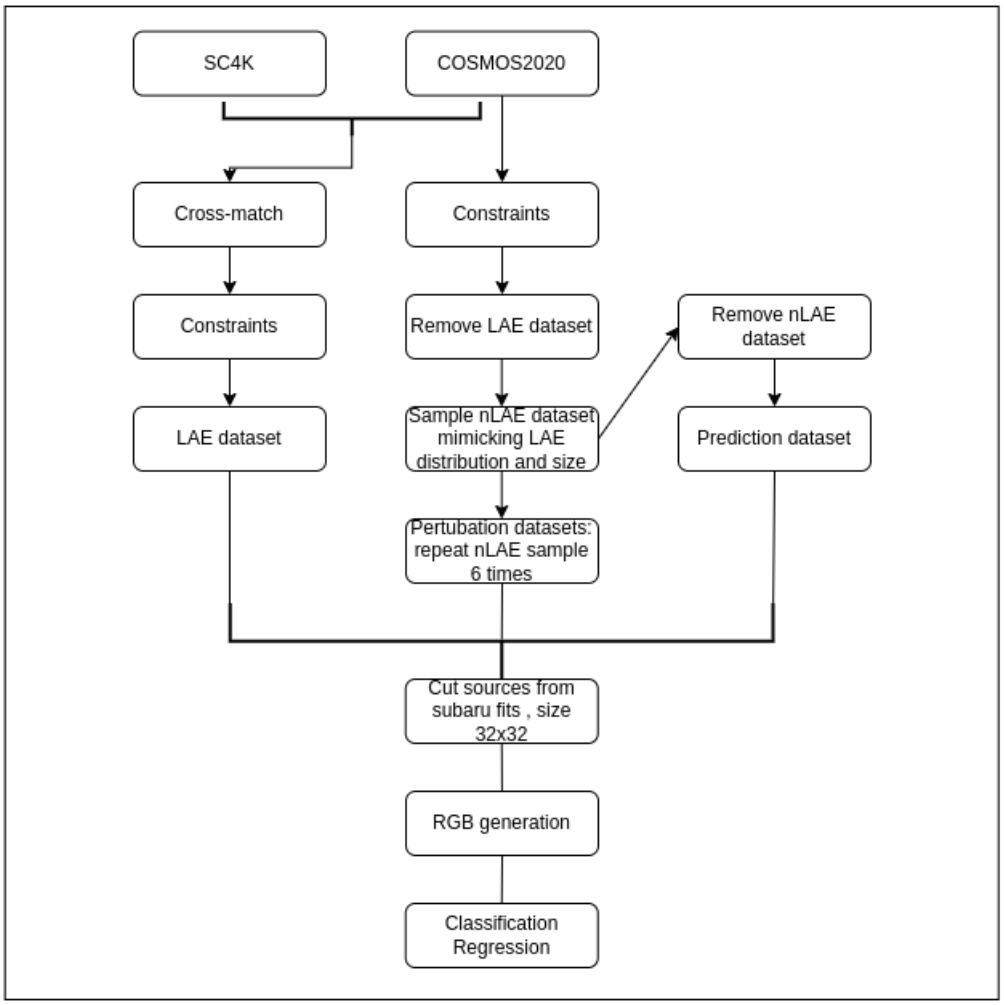


FIGURE 3.2: Flowchart describing the complete pipeline used in this work. Starting from the SC4K survey and COSMOS2020 catalogs, cross-matching and constraint filtering are applied to define the LAE and nLAE datasets. The nLAE dataset is sampled to match the size, redshift, and magnitude distribution of the LAEs, and this sampling is repeated six times to generate perturbed nLAE datasets for robustness analysis. Sources are then cut from Subaru FITS mosaics in the g, r, and i bands with a size of 32×32 pixels. These cutouts are used to generate RGB images in Python, which serve as input for classification and regression models. A separate prediction dataset is built from the remaining COSMOS2020 data, after excluding all LAE and sampled nLAE entries.

3.2.1 Classification

The first step in developing the CNN architecture involved importing the necessary data and libraries using Python. This subsection outlines the architecture, training process, Fine-tune process, and evaluation criteria used for the CNN model and the comparison with other models.

3.2.1.1 Model architecture

The CNN architecture created before tuning is depicted in Figure 3.3, while 3.4 displays the tuned.

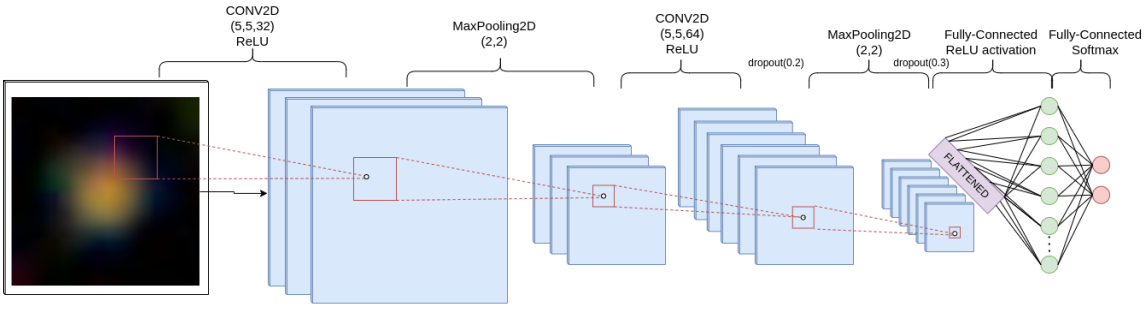


FIGURE 3.3: Initial architecture of the CNN model before hyperparameter tuning. The network consists of two convolutional layers with ReLU activation and max-pooling, followed by a flattening layer and a single fully connected layer with dropout. The final output layer uses softmax activation for binary classification. This configuration was manually defined prior to automated optimization.

The CNN model was designed with an input shape of (32, 32, 3), suitable for processing RGB images constructed from Subaru/HSC broadband filters (g+, r+, i+). The architecture includes the following components:

1. **Convolutional Layers**: Two convolutional layers were used, both with a kernel size of (5, 5). The first layer includes 32 filters, and the second uses 64 filters. The ReLU activation function was chosen for its computational efficiency and ability to introduce non-linearity while mitigating vanishing gradients. The (5, 5) kernel size was selected as a compromise between capturing local spatial features (e.g., compact structures, gradients in brightness) and computational cost. Larger kernels (e.g., 7×7) risk oversmoothing faint features, while smaller kernels (e.g., 3×3) may fail to capture enough structure in low-resolution astrophysical images. The number of filters increases with depth to allow the model to extract a richer hierarchy of features. All convolutional layers include L2 regularization to reduce overfitting [100].

2. **Pooling Layers**: Max pooling layers with a pool size of (2, 2) follow each convolutional layer to reduce the spatial dimensions of the feature maps. The (2, 2) pooling size is widely adopted in astrophysical applications, as it balances dimensionality reduction with the preservation of morphological information. This is particularly relevant when working with small input sizes like 32×32 pixels.

3. **Flatten Layer**: A flatten layer converts the 2D feature maps into a 1D vector, preparing the data for input into the fully connected layers.

4. **Dropout Layers**: A dropout layer with a rate of 0.2 is applied after flattening. Dropout is a common regularization method that randomly disables a fraction of neurons during training, preventing co-adaptation and reducing overfitting. A rate of 0.2 is widely used in CNNs for image classification and regression tasks, offering a balance between regularization and training stability.

5. **Fully Connected Layers**: A single fully connected (dense) layer with 128 units is used after the dropout layer. The number of units was chosen empirically to balance representational capacity and computational cost. Higher dimensional dense layers can easily overfit small image datasets, while fewer units may lack the capacity to model the necessary non-linear relationships. The ReLU activation function is again used here for its simplicity and effectiveness in deep learning models.

6. **Output Layer**: The output layer includes 2 units with a softmax activation function, appropriate for binary classification tasks. In this case, the goal is to distinguish between two galaxy classes (e.g., LAEs and nLAEs), so the softmax outputs the class probabilities.

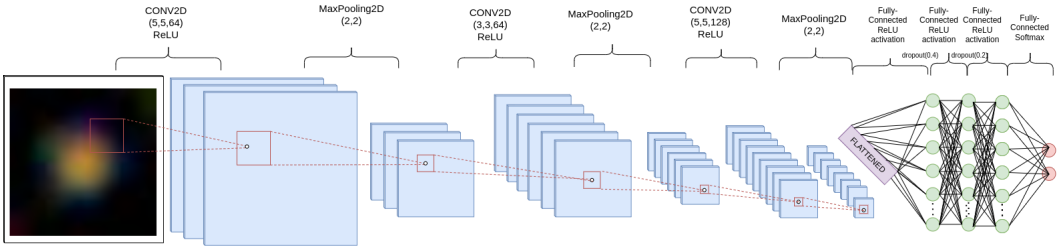


FIGURE 3.4: Final architecture of the CNN model after hyperparameter tuning. The network includes three convolutional blocks with ReLU activations and max-pooling layers, followed by fully connected layers with dropout for regularization. The final dense layer uses a softmax activation for binary classification.

The final architecture resulting from the hyperparameter optimization (Section 3.2.1.2) process consists of a deep convolutional neural network designed to classify RGB galaxy images. The input shape is fixed at (32, 32, 3), corresponding to the RGB composites created from Subaru/HSC filters. The architecture includes the following layers:

- **Convolutional Layers:**
  - First layer: 64 filters, kernel size (5×5), ReLU activation
  - Second layer: 64 filters, kernel size (3×3), ReLU activation
  - Third layer: 128 filters, kernel size (5×5), ReLU activation
- **Pooling Layers:** A `MaxPooling2D` layer with pool size (2×2) is applied after each convolutional layer, reducing spatial resolution while preserving salient features.
- **Flatten Layer:** The output of the final convolutional block is flattened into a one-dimensional vector to be used by the dense layers.
- **Fully Connected Layers:**
  - Dense layer with 256 units and ReLU activation, followed by a `Dropout` layer with a rate of 0.4
  - Dense layer with 128 units and ReLU activation, followed by a `Dropout` layer with a rate of 0.2
  - Dense layer with ReLU activation (unit count as optimized)
  - Final output layer with 2 units and softmax activation for binary classification

This architecture was automatically selected during the hyperparameter tuning phase using Keras Tuner [101], and reflects the best-performing configuration discovered for the LAE classification task.

3.2.1.2 Fine-tune

The model, named **my_CNN**, was compiled and trained using the Keras package [102]. The training process involved multiple epochs, with the test set (10%) used for evaluation. Accuracy was adopted as the primary evaluation metric due to the balanced nature of the dataset. However, additional metrics such as precision, recall, and F1-score were also monitored to detect any unexpected model behavior.

To optimize the model architecture, we employed the Keras Tuner package with the RandomSearch algorithm, using validation accuracy as the objective. The hyperparameters explored included the number of convolutional layers, kernel size, number of filters, L2 regularization strength, number of dense layers, dropout rate, and learning rate. After tuning, the best-performing **my_CNN** configuration was selected and evaluated. The final accuracy and F1-score results are presented in Figure 3.5 and Figure 3.6, respectively.

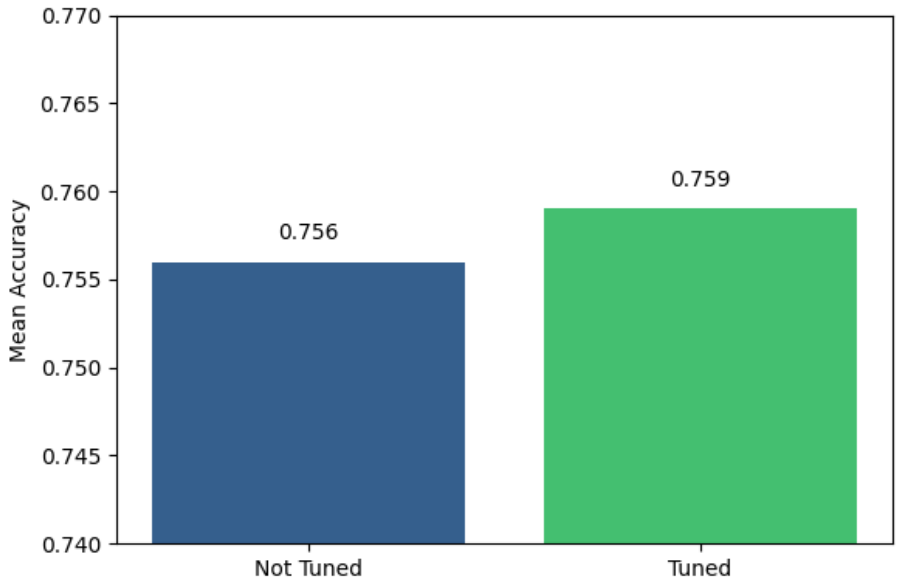


FIGURE 3.5: Accuracy comparison between the CNN with and without tune

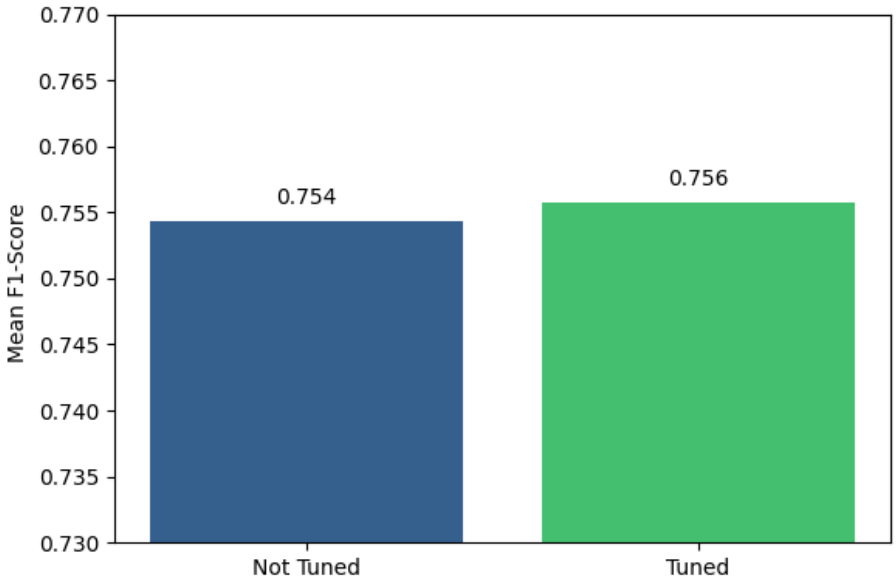


FIGURE 3.6: F1 score comparison between the CNN with and without tune

3.2.1.3 Model comparison

To ensure robustness, **my_CNN** was compared against several well-known CNN architectures: VGG19 [103], Xception [104], ResNet50 [105], DenseNet121 [106], and a visual transform architecture ViT_B/32 [107]. These models were chosen for their diverse architectural strategies: VGG19 is a deep but simple convolutional model with uniform 3×3 filters and around 19 layers; ResNet50 introduces residual connections to enable the training of very deep networks; DenseNet121 densely connects each layer to all preceding ones, promoting feature reuse and parameter efficiency; and ViT_B/32 represents a transformer-based vision model that processes images as sequences of 32×32 patches. Figure 3.7 presents the evaluation results, showing that **my_CNN** and ViT_32b achieved the highest performance metrics.

Figure 3.8 illustrates the architectural complexity in terms of model size. These comparisons supported the choice of **my_CNN Tuned** for the final predictions. Although several models performed similarly in terms of accuracy and F1-score, the final decision prioritized computational efficiency, making the lightweight architecture of **my_CNN** particularly suitable given the available computing resources.

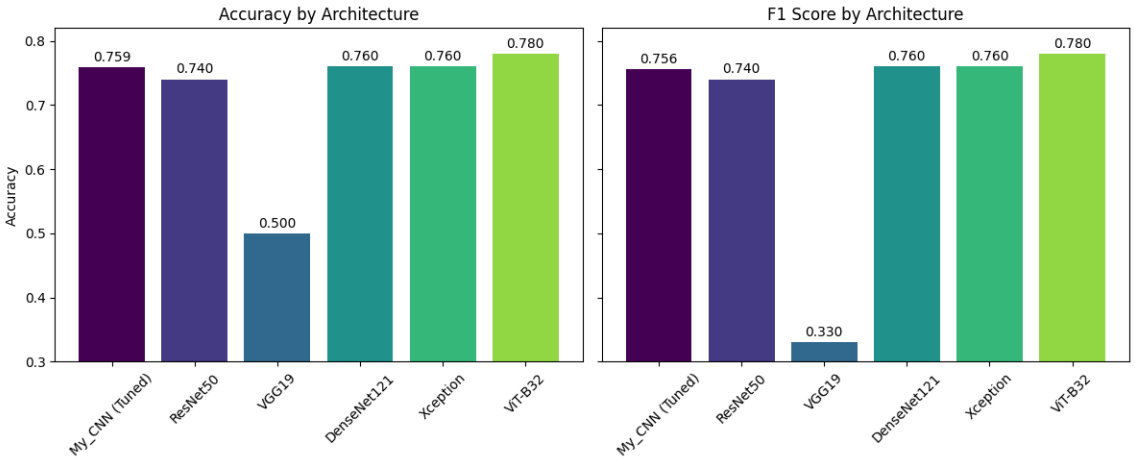


FIGURE 3.7: Comparison of classification accuracy (left) and F1-score (right) for different CNN architectures. ViT_32b achieved the highest accuracy, followed closely by **my_CNN**, DenseNet121, and Xception. For the F1-score, ViT_32b again showed the best performance, with **my_CNN** among the top-performing models.

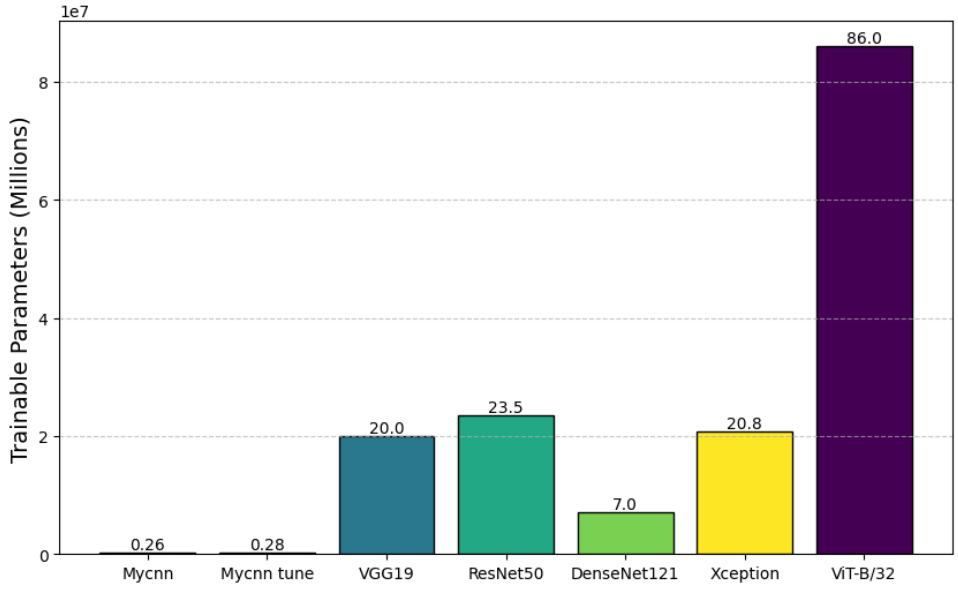


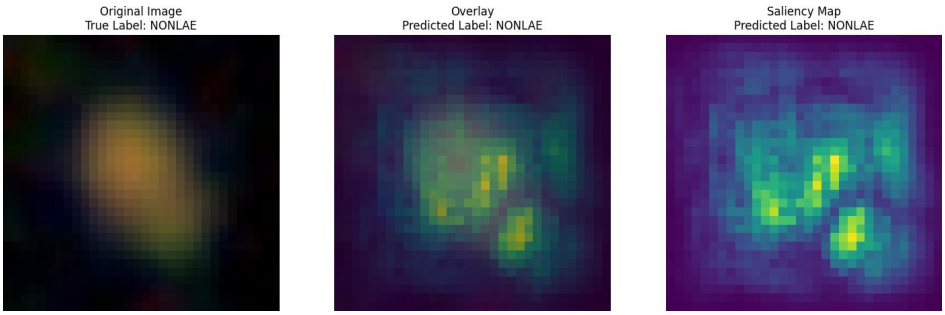
FIGURE 3.8: Total number of trainable parameters per architecture. ResNet50 has the highest number of parameters among the CNNs tested, while **my_CNN** has the lowest, indicating a more lightweight and computationally efficient model.

3.2.1.4 Perturbation analysis and saliency maps

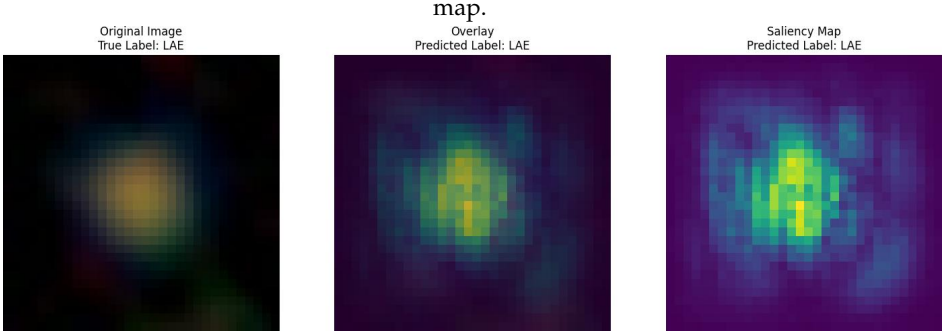
Following model selection, a perturbation analysis was performed using **my_CNN Tuned** to assess the robustness of the classification performance. This analysis involved generating six additional nLAE samples, in addition to the base nLAE sample, resulting in a total of seven nLAE datasets. The number of nLAE datasets was limited to seven due to constraints in the number of available sources within the redshift and i-band magnitude bins. Since the goal was to replicate the distribution of the LAE dataset across these bins, only seven balanced nLAE datasets could be constructed under these conditions. Each of these nLAE datasets was then combined with the same LAE dataset, and the CNN was trained separately for each combination. The resulting performances were compared to evaluate the influence of the different samples on the model's behavior and to confirm the consistency of the predictions.

To further investigate the decision process of the CNN, saliency maps were generated using the SmoothGrad method [108]. Saliency maps are visualizations that highlight the regions of the input images that most strongly influence the model's predictions, offering insight into what parts of the image the network attends to during classification.

Figure 3.9 shows two examples of saliency maps. In each case, the first row (A) corresponds to a source labeled and predicted as nLAE, and the second row (B) corresponds to a source labeled and predicted as LAE. For both examples, the three columns represent: (1) the original RGB image, (2) the image overlaid with saliency, and (3) the isolated saliency map.



(A) Example of a source labeled and predicted as nLAE. From left to right: original RGB image, RGB image with saliency overlay, and isolated saliency map.



(B) Example of a source labeled and predicted as LAE. From left to right: original RGB image, RGB image with saliency overlay, and isolated saliency map.

FIGURE 3.9: Saliency maps generated using the SmoothGrad method. Each row corresponds to a different class (A: nLAE, B: LAE), showing how the model identifies class-relevant regions in the images.

3.2.2 Regression

In this study, the regression approach involves several key steps to prepare and use the data effectively for predicting the astrophysical parameters of LAEs. The process begins with the preparation and labeling of the dataset, followed by loading the data using Keras and Python.

3.2.2.1 Independent CNN models for each target

In this approach, the strategy was to create separate CNNs for each feature. It is illustrated in Figures 3.10a - 3.11. Each CNN was tailored to predict a specific feature independently:

- **Independent $L_{Ly\alpha}$ model** (Figure 3.11): This model focused solely on predicting $L_{Ly\alpha}$, leveraging a specialized architecture.
- **Independent redshift model** (Figure 3.10a): A distinct model was designed to predict the redshift, with an architecture optimized for this task.

- **Independent $\log_{10}(EW_0)$ model** (Figure 3.10b): Another separate model was constructed to predict the LogEW, ensuring that the unique characteristics of this feature were addressed.

This approach yielded good results for redshift and $\log_{10}(EW_0)$, as each CNN could focus on the specific features relevant to its prediction task. However, there was still room for improvement, particularly in the prediction of Lyman-α luminosity. In Section-4.2, the results are presented.

The Tables 3.1, 3.2, and 3.3 provide detailed breakdowns of the individual models' architectures used in this attempt. Each table includes the type of each layer, the input and output shapes, and the number of parameters for each layer.

TABLE 3.1: Describes the structure of the best model for predicting LogEW0, highlighting its layers and the parameter counts, totalizing 191,009 parameters.

Layer (type)	**Output Shape**	**Param #**
conv2d_input (InputLayer)	(None, 32, 32, 3)	0
conv2d (Conv2D)	(None, 30, 30, 32)	896
max_pooling2d (MaxPooling2D)	(None, 15, 15, 32)	0
conv2d_1 (Conv2D)	(None, 13, 13, 32)	9,248
max_pooling2d_1 (MaxPooling2D)	(None, 6, 6, 32)	0
flatten (Flatten)	(None, 1152)	0
dense (Dense)	(None, 128)	147,584
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 256)	33,024
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 1)	257
Total		**191,009**

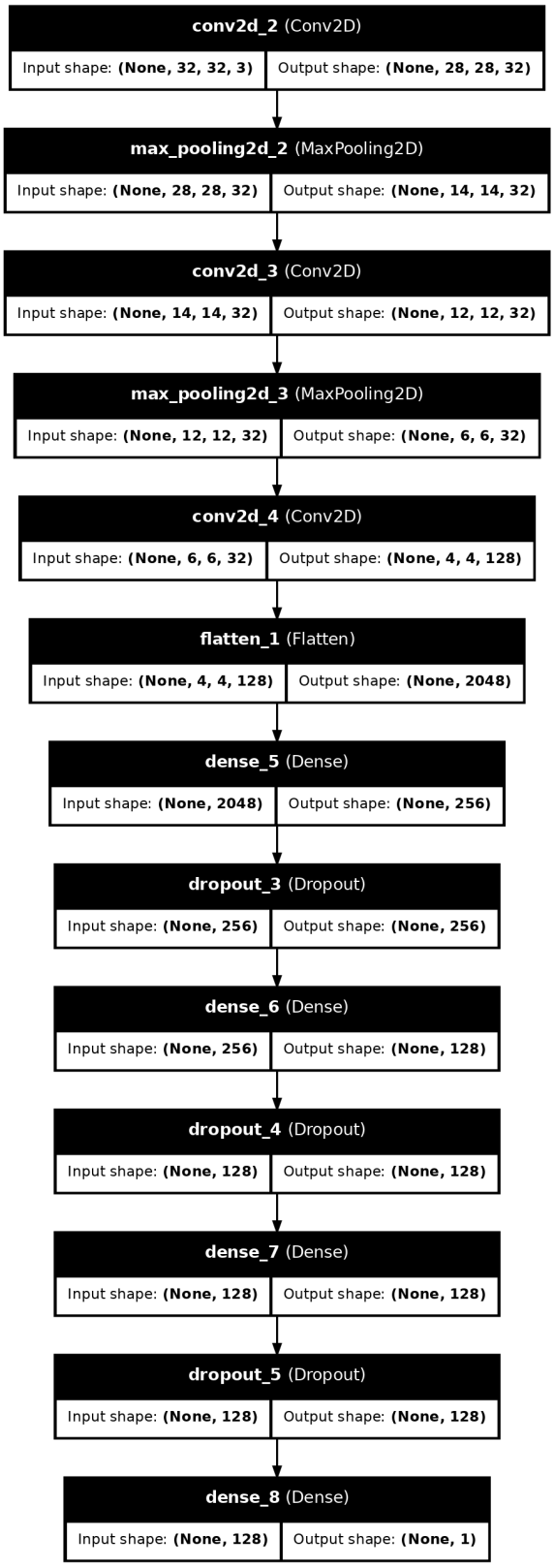
These tables illustrate the tailored architectures for each specific feature, allowing for focused learning and improved prediction accuracy compared to the joint model approach. Each model's parameter count reflects its complexity and capacity to capture the unique characteristics of the feature it was designed to predict.

TABLE 3.2: Summarizes the Lyman-α luminosity model, showing a total of 95,299 parameters.

Layer (type)	**Output Shape**	**Param #**
conv2d (Conv2D)	(None, 28, 28, 32)	2,432
max_pooling2d (MaxPooling2D)	(None, 14, 14, 32)	0
conv2d_1 (Conv2D)	(None, 12, 12, 32)	9,248
max_pooling2d_1 (MaxPooling2D)	(None, 6, 6, 32)	0
conv2d_2 (Conv2D)	(None, 2, 2, 32)	25,632
max_pooling2d_2 (MaxPooling2D)	(None, 1, 1, 32)	0
flatten (Flatten)	(None, 32)	0
dense (Dense)	(None, 256)	8,448
dropout (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 128)	32,896
dropout_1 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 128)	16,512
dropout_2 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 1)	129
Total		**95,299**

TABLE 3.3: Outlines the redshift model, with a total of 95,297 parameters.

Layer (type)	**Output Shape**	**Param #**
conv2d (Conv2D)	(None, 28, 28, 32)	2,432
max_pooling2d (MaxPooling2D)	(None, 14, 14, 32)	0
conv2d_1 (Conv2D)	(None, 12, 12, 32)	9,248
max_pooling2d_1 (MaxPooling2D)	(None, 6, 6, 32)	0
conv2d_2 (Conv2D)	(None, 2, 2, 32)	25,632
max_pooling2d_2 (MaxPooling2D)	(None, 1, 1, 32)	0
flatten (Flatten)	(None, 32)	0
dense (Dense)	(None, 256)	8,448
dropout (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 128)	32,896
dropout_1 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 128)	16,512
dropout_2 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 1)	129
Total		**95,297**



(A) Independent Redshift model

(B) Independent $\log_{10}(EW_0)$ model

FIGURE 3.10: Architectures of the CNN models independently trained for redshift and $\log_{10}(EW_0)$ regression.

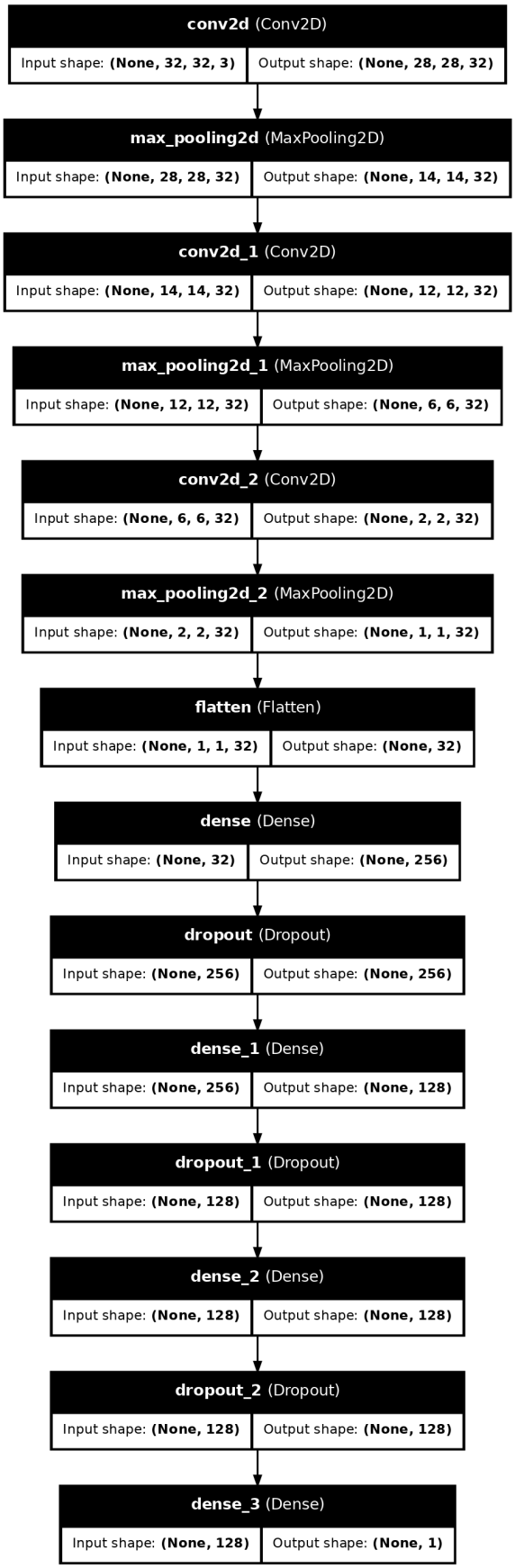


FIGURE 3.11: Architecture of the CNN model independently trained for Lyman-alpha luminosity ($L_{Ly\alpha}$) regression.

3.2.2.2 Chained regression using redshift predictions as auxiliary input for luminosity

This involved a more refined approach, integrating predictions from the redsfhit individual model into subsequent models. This approach is depicted in Figure 3.12. The process involved several key steps:

1. **Redshift prediction**: First, trained a model to predict the redshift independently, with the same model used already for redshift.

2. **Chained prediction for Lyman-α luminosity**: The predictions from the redshift model were then used as additional input features for a second CNN designed to predict the Lyman-α luminosity. This integration occurred after the convolutional layers but before the fully connected layers; the input of redshift is not given initially to the model, only after it passes through the convolutional layers.

3. $\log_{10}(EW_0)$ **prediction**: The structure from the independent models was retained for predicting $\log_{10}(EW_0)$, ensuring consistency and leveraging the previously identified strengths.

This approach used the best results of the independent approach and joined together with a new one for the Lyman-α luminosity, with the incorporation of the redshift predictions as input to the Lyman-α luminosity prediction model.

The Table 3.4 provides a detailed breakdown of the model architecture used for predicting Lyman-α luminosity. The structure leverages the predictions from the redshift model, concatenating the output with the image features to improve prediction accuracy. The total number of parameters in this model is 441,377. For the $\log_{10}(EW_0)$ and redshift predictions, we used the same models described in the independent CNN models, detailed in Tables 3.1 and 3.3, respectively.

TABLE 3.4: Structure of the Model Lyman Ensemble

Layer (type)	**Output Shape**	**Param #**
image_input (InputLayer)	(None, 32, 32, 3)	0
conv2d_12 (Conv2D)	(None, 30, 30, 32)	896
max_pooling2d_10 (MaxPooling2D)	(None, 15, 15, 32)	0
conv2d_13 (Conv2D)	(None, 13, 13, 32)	9248
max_pooling2d_11 (MaxPooling2D)	(None, 6, 6, 32)	0
conv2d_14 (Conv2D)	(None, 4, 4, 128)	36992
max_pooling2d_12 (MaxPooling2D)	(None, 2, 2, 128)	0
flatten_5 (Flatten)	(None, 512)	0
dropout_14 (Dropout)	(None, 512)	0
redshift_input (InputLayer)	(None, 1)	0
concatenate (Concatenate)	(None, 513)	0
dense_21 (Dense)	(None, 512)	262656
dropout_15 (Dropout)	(None, 512)	0
dense_22 (Dense)	(None, 256)	131328
dense_23 (Dense)	(None, 1)	257
Total		**441,377**

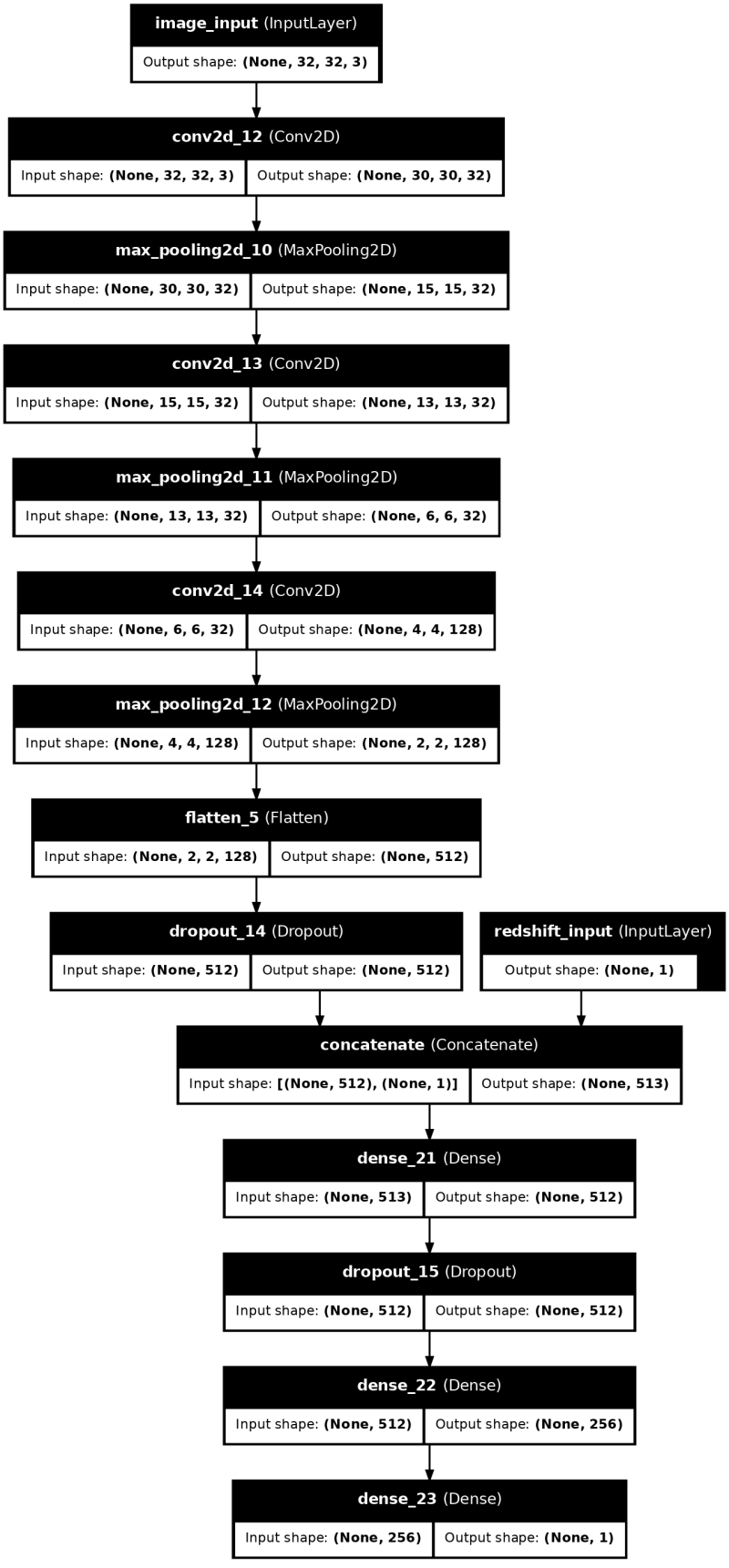


FIGURE 3.12: Chained regression utilizing the redshift as input to help predicting $L_{Ly\alpha}$.

3.2.2.3 Joint multi-target regression with a single CNN

The best results were found when it was aimed to predict the redshift, Lyman-α luminosity, and equivalent width ($\log_{10}(EW_0)$) simultaneously using a single CNN. The architecture for this approach is shown in Figure 3.13. In this model, the three target features were input together, and the network was trained to predict them concurrently. However, this approach faced significant challenges for the implementation. The input of the features values occurred after the convolutional layers.

TABLE 3.5: Structure of the Model Together First Attempt

Layer (type)	**Output Shape**	**Param #**
input_layer_3 (InputLayer)	(None, 32, 32, 3)	0
conv2d_7 (Conv2D)	(None, 28, 28, 32)	896
max_pooling2d_6 (MaxPooling2D)	(None, 14, 14, 32)	0
conv2d_8 (Conv2D)	(None, 10, 10, 64)	18496
max_pooling2d_7 (MaxPooling2D)	(None, 5, 5, 64)	0
flatten_3 (Flatten)	(None, 1600)	0
dense_12 (Dense)	(None, 128)	204928
dropout_8 (Dropout)	(None, 128)	0
dense_13 (Dense)	(None, 128)	16512
dropout_9 (Dropout)	(None, 128)	0
dense_14 (Dense)	(None, 256)	33024
dropout_10 (Dropout)	(None, 256)	0
dense_15 (Dense)	(None, 256)	65792
EWlog (Dense)	(None, 1)	257
LyAlpha (Dense)	(None, 1)	257
dense_16 (Dense)	(None, 256)	65792
redshift (Dense)	(None, 1)	257
Total		**387,211**

The table 3.5 provides a detailed breakdown of the model's architecture, including the type of each layer, the output shape, and the number of parameters for each layer. The total number of parameters in this model is 387,211. This significant number of parameters reflects the complexity of the model and highlights the challenges faced in training a model to predict three different astrophysical features simultaneously.

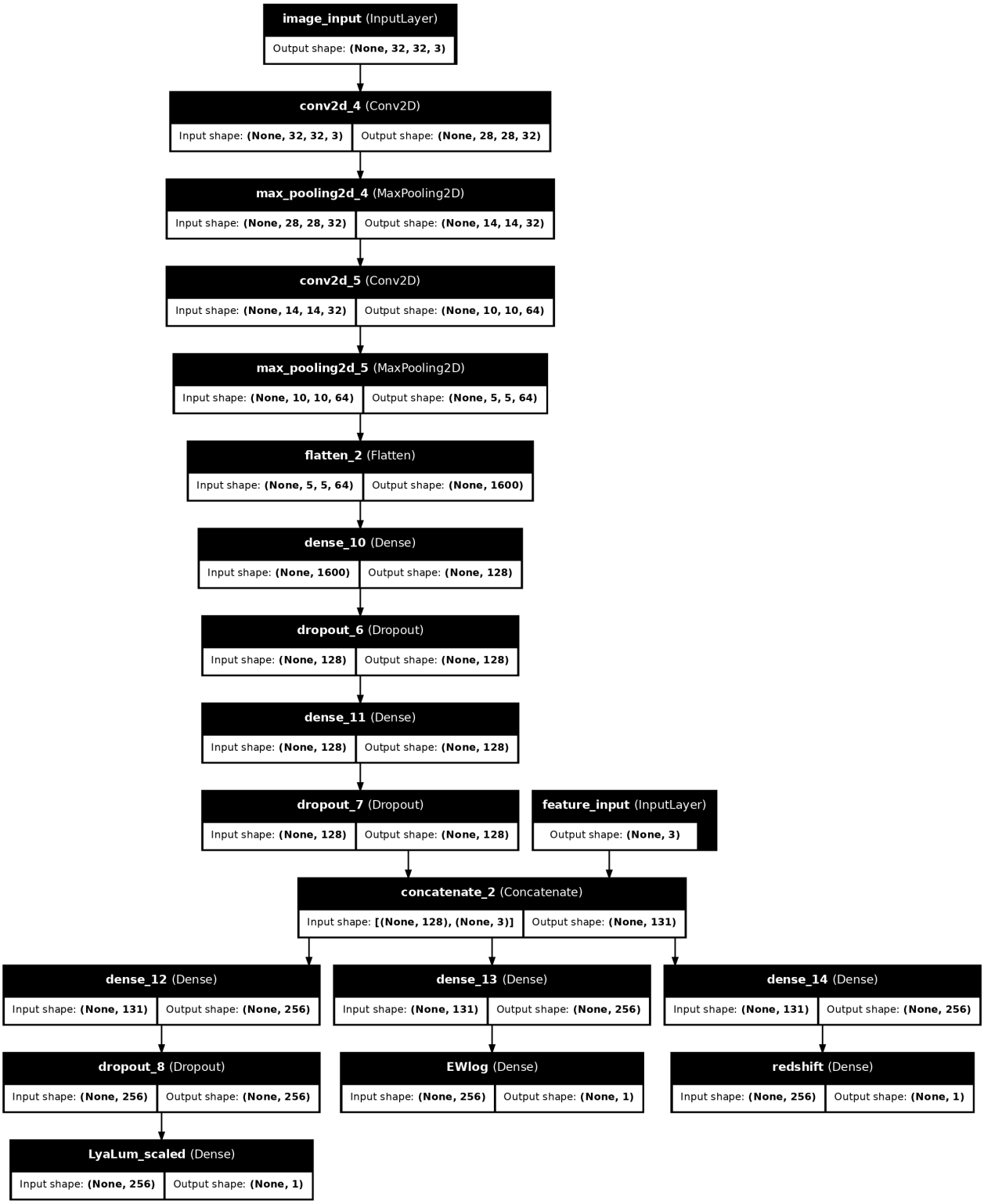


FIGURE 3.13: Structure of the CNN used in the joint multi-target regression with a single CNN.

3.2.2.4 Performance evaluation

To evaluate the performance of the regression models, we focused on the MAE and MSE. These metrics provided a comprehensive assessment of the models' accuracy and robustness, guiding our iterative improvements and validating the effectiveness of our final approach.

Chapter 4

Results

This section presents the results obtained from the classification task, along with a detailed analysis of the outcomes from the three approaches applied to the regression models. The performance of these methods is evaluated using their respective metrics, providing a comprehensive view of their effectiveness. Additionally, potential factors influencing these results are discussed, emphasizing the strengths and limitations observed throughout the experiments.

4.1 Classification results

The performance comparison between models revealed minimal differences in accuracy, as shown in Figure 3.7. However, their computational requirements varied significantly. Ultimately, the best results were obtained using the custom CNN architecture developed specifically for this task, referred to as **my_CNN Tuned**, which achieved an accuracy of 75.9% on the test set, as illustrated in Figure 3.5.

The model was then applied to the prediction dataset, which contains 191,826 sources. Using a threshold of 0.5, where 0 corresponds to LAE and 1 to nLAE, a total of 44,295 sources were classified as LAEs. These are LAEs candidate that exhibit similar features, regarding redshift and i-band magnitude, to those in the SC4K catalog [1].

Among the predicted LAEs, some sources had already been spectroscopically confirmed in prior studies, particularly in the HETDEX survey [109]. HETDEX (Hobby-Eberly Telescope Dark Energy Experiment) is a large-scale integral field spectroscopic survey targeting the spatial distribution of Lyα-emitting galaxies over a wide area of 540 deg^2. It aims to constrain cosmological parameters by measuring the Hubble expansion

rate and angular diameter distance in the redshift range $1.88 < z < 3.52$. To assess the model's behavior in such cases, a comparison was conducted using the 45 sources in common between our prediction dataset and HETDEX. The resulting confusion matrix is presented in Figure 4.1.

In this evaluation, the CNN model "correctly" identified 21 true LAEs and 2 true nLAEs. However, it also misclassified 18 LAEs as nLAEs (false negatives) and "incorrectly" labeled 4 nLAEs as LAEs (false positives), resulting in an LAE classification precision of 84%. This indicates that while the model is relatively conservative and precise when assigning the LAE label, it may fail to capture some true emitters. It is also important to note that the HETDEX dataset has its own limitations, but this cross-comparison provides a valuable validation step.

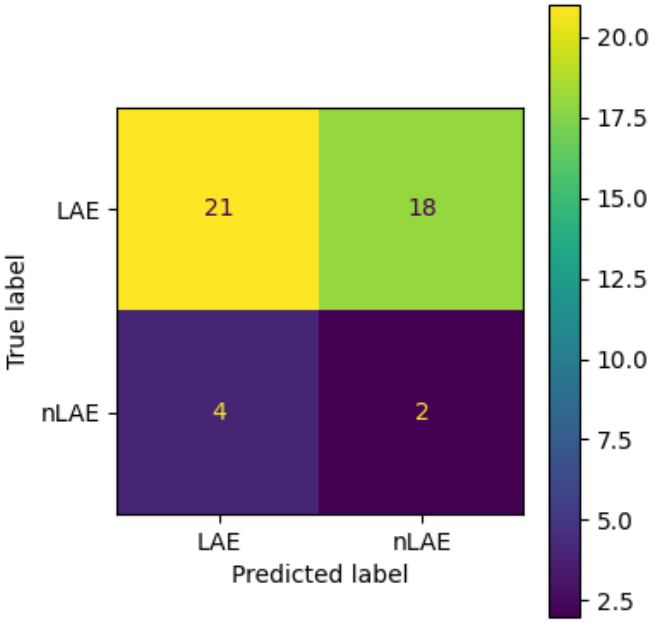


FIGURE 4.1: Confusion Matrix of the sources spectroscopically confirmed by HETDEX [109]

Table 4.1 lists the top 15 most confidently predicted LAEs from the classification model. Each entry includes RA, Dec, the photometric redshift (z_{phot}), i-band magnitude, and the predicted probability.

Table 4.2 shows the top 15 predictions from the model that also have spectroscopic redshifts available from HETDEX. This comparison helps identify which predicted LAEs are reinforced by independent observations.

TABLE 4.1: Top 15 most confidently predicted LAEs by the classification model. The table includes each source's coordinates (RA and Dec), the photometric redshift from COSMOS2020 (z_{phot}), the i-band magnitude, and the prediction probability (0: LAE, 1: nLAE). Complete table available at: [https://github.com/Onirb/Tese/blob/main/Tables/tabela_completa_com_probabilidades.csv].

RA (deg)	**Dec (deg)**	z_{phot}	i_band_mag	prob
149.6052	2.5349	2.0841	25.9542	0.0003
149.8030	1.6994	2.2579	26.5433	0.0005
150.7421	2.7697	2.8810	26.2006	0.0016
150.7462	2.7659	2.4634	27.4887	0.0022
149.7642	2.5542	2.2889	24.9029	0.0035
149.5119	2.1461	2.5392	24.3433	0.0051
150.1451	2.6124	2.1346	25.6657	0.0063
149.9902	1.9977	2.5024	25.6428	0.0082
150.7423	2.2091	2.0528	25.9492	0.0082
150.3735	2.2066	2.7508	25.6328	0.0087
150.6682	1.7012	2.7230	26.1638	0.0090
150.7449	2.7667	2.1973	25.9406	0.0095
149.6809	1.9943	2.6065	26.2841	0.0100
150.5346	2.6408	2.6965	24.9943	0.0110
149.5775	1.6949	2.3068	25.8649	0.0112

TABLE 4.2: Top 15 predicted LAEs with spectroscopic matches from HETDEX. The table includes RA, Dec, COSMOS2020 photometric redshift (z_{phot}), HETDEX spectroscopic redshift (z_{HETDEX}), and model prediction probability(0: LAE, 1: nLAE).Complete table available at: https://github.com/Onirb/Tese/blob/main/Tables/tabela_HETDEX_com_probabilidades.csv.

RA (deg)	**Dec (deg)**	z_{phot}	z_{HETDEX}	prob
150.2482	2.2772	2.7034	2.8796	0.0440
150.1269	2.3682	2.7841	2.6753	0.1210
150.1157	1.9195	3.2152	3.3174	0.1248
150.2782	2.2524	2.8423	2.6113	0.1254
150.2658	2.3661	2.3766	2.5523	0.1337
150.1989	2.2578	2.6410	2.6106	0.1412
150.1359	2.3131	2.4953	2.6007	0.1616
150.2562	2.3822	2.4029	2.3188	0.1668
150.2027	2.2003	2.5988	2.6330	0.1893
150.1205	2.1923	2.7757	2.6916	0.1926
150.1211	2.2354	2.3764	2.4348	0.1974
149.8765	1.9346	2.1508	2.3867	0.2458
150.2009	2.2258	2.8700	2.4815	0.2594
150.0683	2.2165	2.2655	3.2560	0.3257
150.2540	2.2338	2.0128	0.0000	0.3451

4.2 Regression results

The regression approach involved multiple attempts to optimize the prediction of astrophysical parameters of LAEs. These attempts included joint regression with single CNN,

independent models for each feature, and an integrated approach using redshift predictions as inputs to luminosity.

4.2.1 Independent CNN models for each target

Independent CNNs were developed for each individual target: redshift, equivalent width ($\log_{10}(EW_0)$), and Lyman-α luminosity. This modular approach led to good performance metrics across most outputs, as each model could specialize in learning the specific patterns relevant to its target. The Figures 4.2-4.7 illustrate the results for each of the three models.

The redshift CNN achieved a test loss (MAE) of 0.263 and a Root Mean Squared Error (RMSE) of 0.352. By analyzing the training evolution shown in Figure 4.2, one might infer that extending the number of epochs could lead to further improvements. However, another interpretation is that the network architecture may lack sufficient complexity or information to enhance performance beyond this point. Despite these limitations, the model predictions remain close to the true redshift values, as illustrated in the histogram of Figure 4.3.

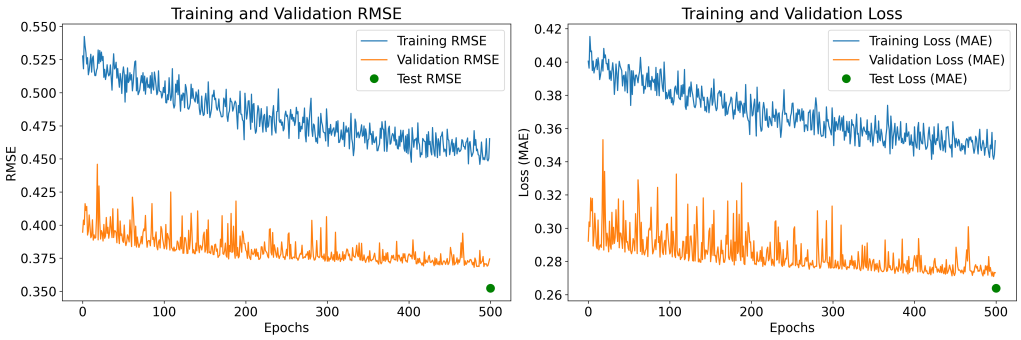


FIGURE 4.2: Performance of the redshift CNN trained independently. The plot shows the training and validation MAE over epochs.

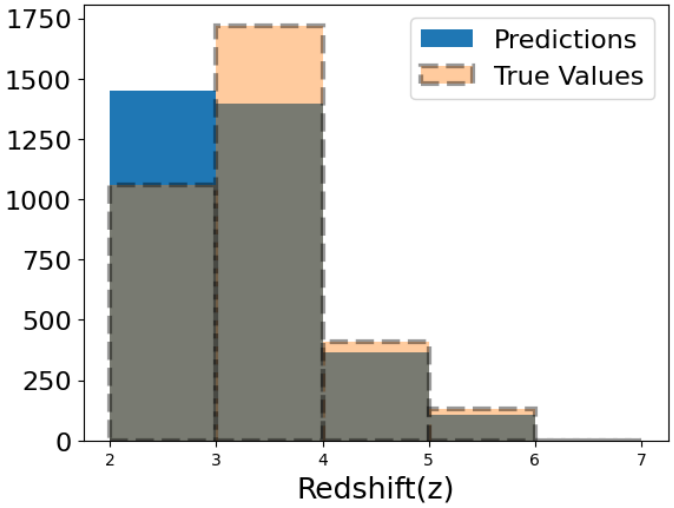


FIGURE 4.3: Histogram of the redshift predictions generated by the independent CNN model.

For the $\log_{10}(EW_0)$, the model achieved a test loss (MAE) of 0.266 and an RMSE of 0.315, indicating strong predictive performance for equivalent width. As shown in Figure 4.4, the training, validation, and test metrics are already closely aligned, suggesting that further improvements may only be possible by modifying the architecture or providing additional input features. Furthermore, Figure 4.5 demonstrates that the predicted values are well aligned with the true labels, reinforcing the model's reliability.

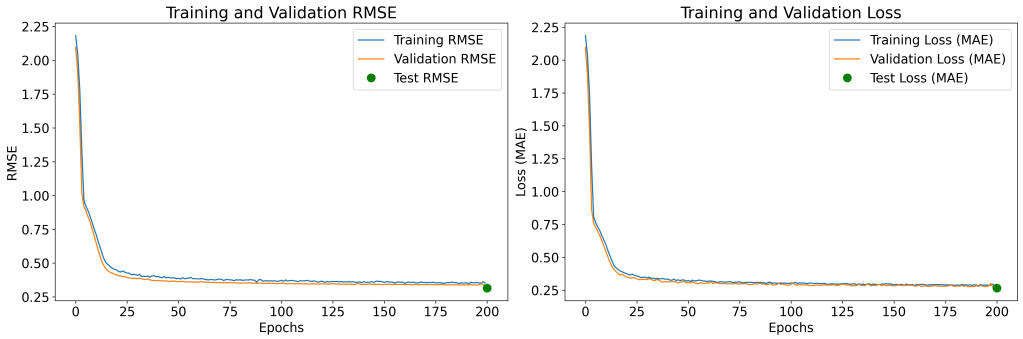


FIGURE 4.4: Performance of the $\log_{10}(EW_0)$ model in the independet CNN. The figure presents the MAE evolution during training and validation.

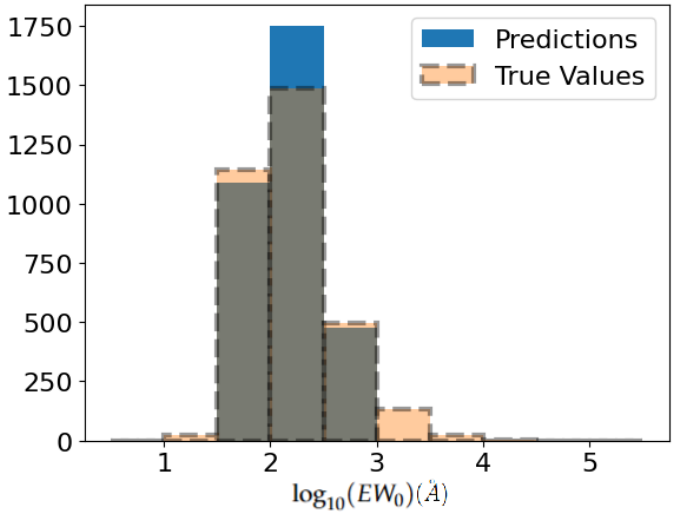


FIGURE 4.5: Histogram of the predicted equivalent width values ($\log_{10}(EW_0)$) from the independent CNN.

The Lyman-α luminosity model showed room for improvement, with a test loss (MAE) of 0.622 and an RMSE of 0.714. As shown in Figure 4.6, the model presents consistently lower training error compared to validation and test metrics, indicating signs of overfitting. This pattern appears early in training and persists throughout, suggesting that while the CNN is able to learn certain features, it fails to generalize effectively to unseen data. Consequently, this model yields the weakest performance among the three regressors. This is further reflected in the distribution shown in Figure 4.7, where the predicted values show larger deviations from the true luminosity values.

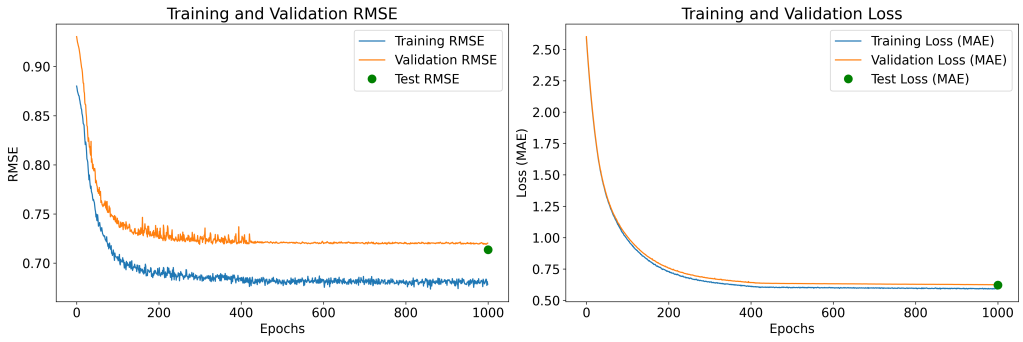


FIGURE 4.6: Performance of the Lyman-alpha luminosity model in independent CNN. The training and validation MAE curves are shown.

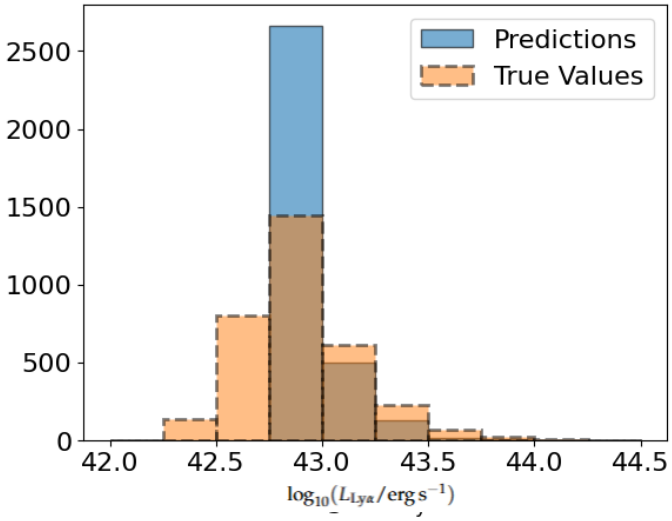


FIGURE 4.7: Histogram of the predicted Lyman-α luminosity values ($L_{Ly\alpha}$) from the independent CNN in.

4.2.2 Chained regression using redshift predictions as auxiliary input

In this strategy, the CNN trained to predict Lyman-α luminosity received not only image and tabular data, but also the redshift predictions generated by the corresponding independent redshift model from the previous attempt. This chained or ensemble approach aimed to leverage the relatively strong redshift predictions to improve the more challenging task of Lyman-α luminosity estimation. The architecture used for this model remains the same as in the independent setup, ensuring consistency and isolating the effect of the added input feature.

The inclusion of redshift as an auxiliary input led to a test loss (MAE) of 0.564 and an RMSE of 0.699, representing a moderate improvement over the independent Lyman-α model discussed in the previous section. As shown in Figure 4.8, the training error continues to decrease while the validation and test errors plateau, suggesting that overfitting remains a concern, likely due to excessive training epochs rather than model capacity. Additionally, Figure 4.9 shows that, although predictions are better aligned with the true luminosity values compared to the previous model, there is still noticeable deviation. This highlights that, despite improvement, Lyman-α luminosity remains the most difficult parameter to predict, and further enhancements may require architectural changes or the incorporation of more discriminative features.

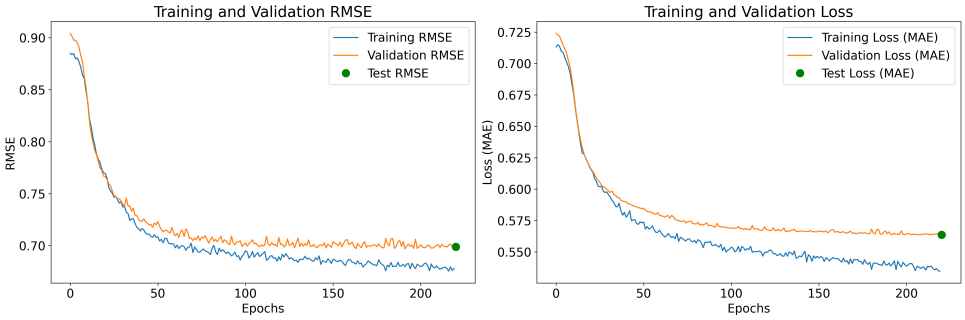


FIGURE 4.8: Performance of the Lyman-α luminosity CNN model trained using redshift predictions as auxiliary input. The plot shows training and validation MAE across epochs.

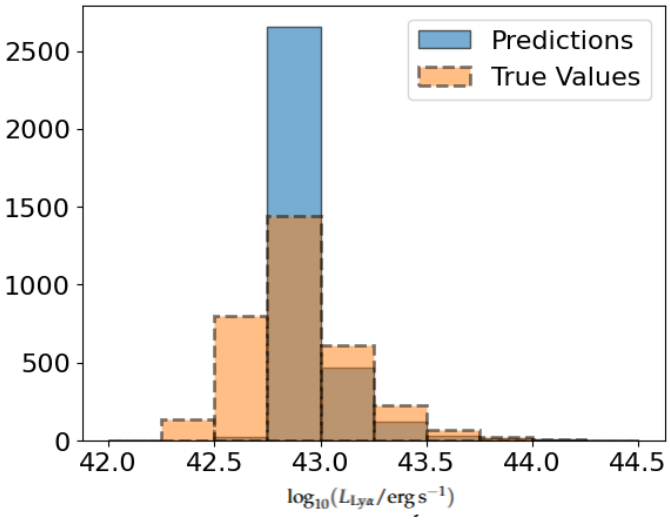


FIGURE 4.9: Histogram of the predicted Lyman-α luminosities ($L_{Ly\alpha}$) produced using redshift predictions as auxiliary input to the regression model.

4.2.3 Joint multi-target regression with a single CNN

The final approach involved a single CNN trained to simultaneously predict redshift, Lyman-α luminosity ($L_{Ly\alpha}$), and logarithmic equivalent width ($\log_{10}(EW_0)$), as illustrated in Figure 4.10. Despite the architectural complexity required to jointly model three different outputs, this strategy produced the best overall results among all experiments, particularly showing a significant improvement in the $L_{Ly\alpha}$ predictions. The mean MAE across all outputs was 0.032, although individual MAEs and RMSEs were not recovered during training due to technical limitations.

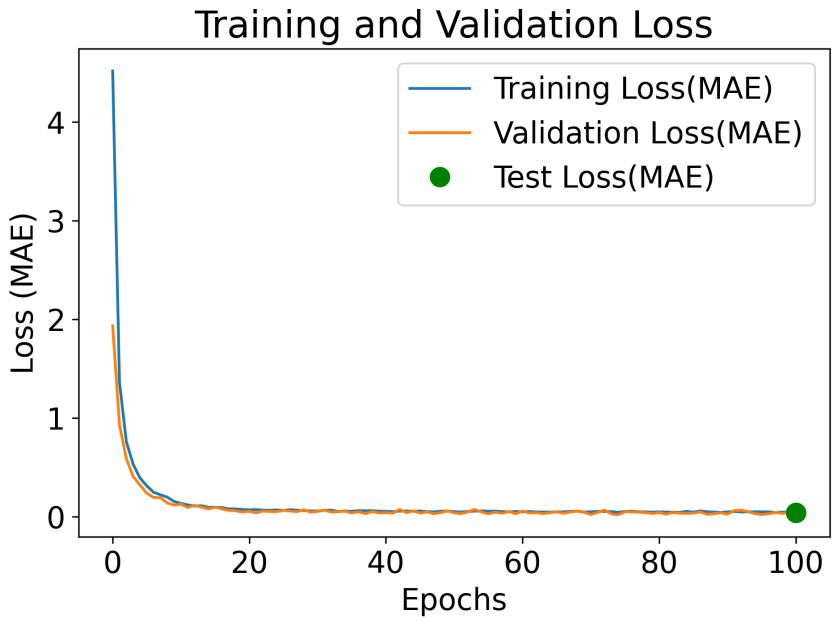


FIGURE 4.10: Performance of the CNN trained to jointly predict all target features (Lyman-α luminosity, equivalent width, and redshift) using both imaging and tabular inputs. The y-axis represents the MAE, and the x-axis shows the training progress over epochs.

To better understand the model's behavior, Figures 4.11–4.13 show the histograms of the predicted values for each output variable. The distributions suggest a better alignment between the predicted and true values across all three features, when compared to the previous independent and chained models. These visual results reinforce the conclusion that the joint multi-target regression architecture achieved the most balanced and robust performance.

The third attempt proved to be the most successful, as it combined the strengths of the previous approaches. This ensemble strategy allowed the CNN to leverage the redshift predictions as additional inputs, enhancing the accuracy of Lyman-α luminosity predictions.

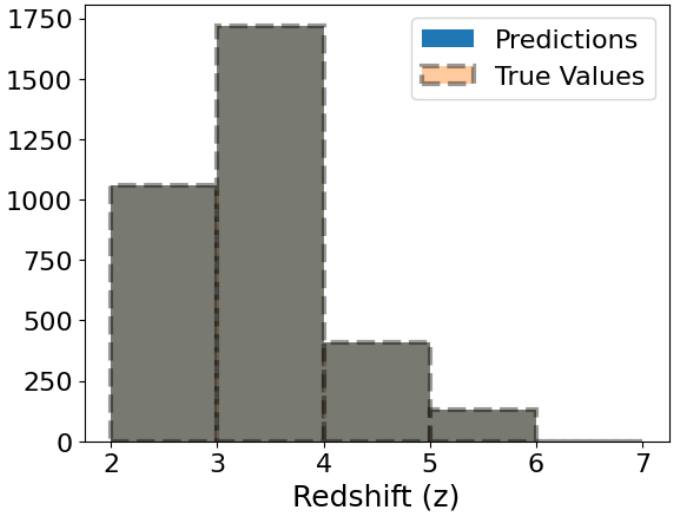


FIGURE 4.11: Histogram of the redshift predictions obtained from the joint multi-target CNN regression model. The distribution reflects the model's output across the test or prediction set.

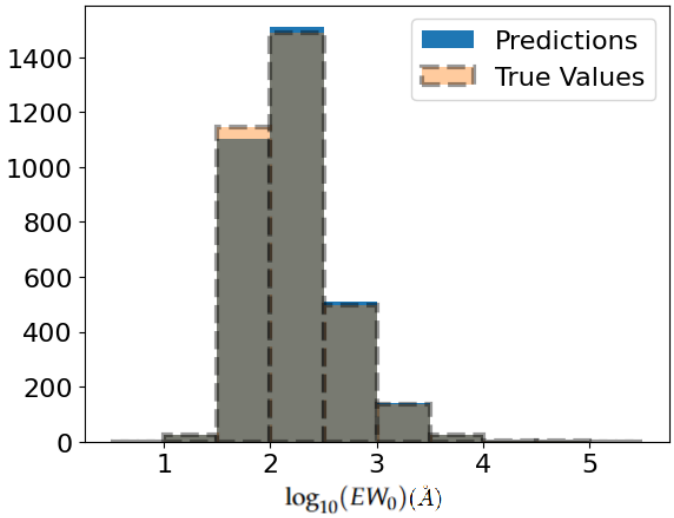


FIGURE 4.12: Histogram of the predicted values of the logarithmic equivalent width ($\log_{10}(EW_0)$) from the joint multi-target CNN regression.

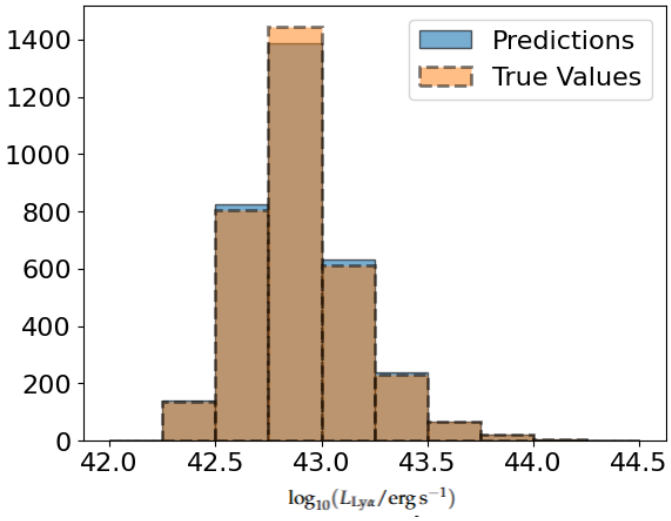


FIGURE 4.13: Histogram of the predicted scaled Lyman-α luminosities ($L_{Ly\alpha}$) produced by the joint multi-target regression model.

4.2.4 Summary and comparison of regression strategies

Three regression strategies were explored to predict the redshift, equivalent width, and Lyman-α luminosity of LAEs using CNNs trained on image and tabular data. The independent models achieved strong results for redshift and equivalent width, but struggled with $L_{Ly\alpha}$. The chained approach, which incorporated redshift predictions into the luminosity model, showed modest improvements, indicating that prior information is valuable. Finally, the joint multi-target CNN achieved the most robust and consistent performance across all outputs. While individual MAE values could not be extracted for this last model, both training curves and histogram comparisons revealed more accurate and stable predictions, particularly for the Lyman-α luminosity. These results suggest that sharing learned representations across related astrophysical features may help overcome limitations observed in single-output models.

To illustrate the outcome of the regression process, Table 4.3 presents a random selection of 15 entries from the prediction dataset, showing the Joint multi-target model's outputs alongside relevant catalog information. Each row corresponds to a galaxy from the COSMOS2020 catalog, including its right ascension (RA), declination (Dec), and original photometric redshift estimate (z_{phot}). The predicted quantities, redshift, Lyman-α luminosity ($\log_{10}(L_{Ly\alpha})$), and equivalent width ($\log_{10}(EW_0)$), were generated using the final joint multi-target CNN regression model, which was trained to estimate all three properties simultaneously. A link to the full table will be available in the appendix.

TABLE 4.3: Random 15 predictions generated by the joint multi-target CNN regression model. The table includes each source's coordinates (RA and Dec), the photometric redshift from COSMOS2020 (z_{phot}), and both the true and predicted values for redshift, Lyman-α luminosity, and equivalent width, complete Table available in [https://github.com/Onirb/Tese/blob/main/Tables/Regression_Predictions.csv].

RA (deg)	**Dec** (deg)	z_{phot}	z_{true}	z_{pred}	$\log_{10}(L_{Ly\alpha}/erg\ s^{-1})$ (true)	$\log_{10}(L_{Ly\alpha}/erg\ s^{-1})$ (pred)	$\log_{10}(EW)$ ($\AA$) (true)	$\log_{10}(EW)$ ($\AA$) (pred)
150.1229	2.2246	2.4845	2.5100	2.5251	42.4935	42.4861	1.8261	1.8300
150.0564	2.6180	3.0475	3.4000	3.4086	42.8525	42.8541	3.0590	3.0689
149.7870	2.1277	2.7850	3.0100	3.0205	42.8055	42.8028	1.9295	1.9378
150.7280	1.9997	0.1497	3.4000	3.4127	42.7323	42.7260	1.7639	1.7727
149.9117	2.0689	0.1666	4.4800	4.4861	43.2836	43.2894	1.7750	1.7880
150.2751	1.6679	2.3983	2.5100	2.5236	42.5337	42.5264	2.3088	2.3145
150.1175	2.2315	2.9241	3.0100	3.0222	42.6930	42.6864	1.9436	1.9517
149.8691	1.7411	4.7405	4.8600	4.8775	42.8049	42.8028	1.5887	1.5971
150.2873	2.7654	2.7743	2.8100	2.8202	42.7626	42.7580	1.9960	2.0044
149.4906	1.8574	3.2314	3.4000	3.4096	42.9485	42.9523	1.8045	1.8128
150.5235	2.2009	2.9960	3.1700	3.1819	42.7859	42.7817	1.7984	1.8071
150.1118	2.0467	3.2046	3.4000	3.4120	42.8114	42.8088	1.7621	1.7712
149.7919	2.7342	4.7040	4.7700	4.7786	43.2523	43.2592	2.4005	2.4145
149.5655	1.6457	2.5112	2.8100	2.8181	42.8258	42.8243	2.3867	2.3946
149.8819	1.7261	0.8306	5.7000	5.7171	42.9270	42.9306	3.0019	3.0147

Chapter 5

Discussion

This chapter discusses the interpretation of findings, model limitations, and their scientific implications.

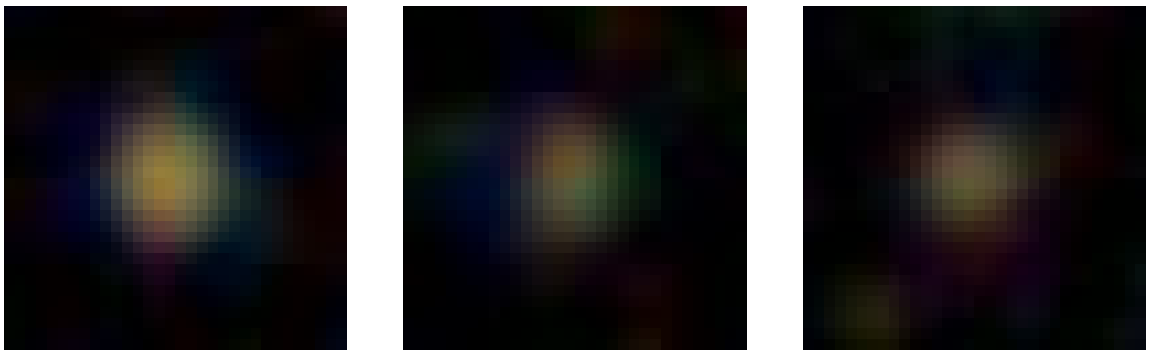
5.1 Catastrophic failures

In the analysis of the classification results, the top 5 catastrophic failures for each category were identified in the first dataset and subsequent datasets. These errors represent instances in which the model made incorrect predictions with the highest confidence, making them critical points for evaluation. Since we used 7 models, generated with the same architecture but varying the nLAE dataset, we obtained 35 catastrophic errors for each class. This analysis was performed over the test split of each dataset, enabling consistent comparison with the original values.

The most significant classification errors for nLAEs in the first dataset are detailed in Table 5.1. These errors were identified by their high confidence deviation, calculated as the absolute difference between the predicted probability and the true label. Ideally, this value should be 0.0 for a perfect LAE prediction and 1.0 for a perfect nLAE prediction. Table 5.5 displays more information for this sources.

TABLE 5.1: Top 5 catastrophic errors with original class nLAE and predicted class LAE, obtained from My_CNN after tune, displaying the COSMOS2020 ID, the predicted probability, predicted class, and original class.

ID	Predicted probability	Predicted class	Original class
449920	0.0027	LAE (0.0)	nLAE (1.0)
1088508	0.008	LAE (0.0)	nLAE (1.0)
1429856	0.0135	LAE (0.0)	nLAE (1.0)
238435	0.0182	LAE (0.0)	nLAE (1.0)
646115	0.0296	LAE (0.0)	nLAE (1.0)



(A) ID: 449920 (B) ID: 1088508 (C) ID: 1429856

(D) ID: 238435 (E) ID: 646115

FIGURE 5.1: RGB images of the Top catastrophic errors for nLAEs. Each subfigure represents a significant classification error, their true label is nLAE and were predicted as LAE.

All five sources appear to be central galaxies with compact morphologies. Additionally, some exhibit blue central emission, which may have contributed to their misclassification. This is because compactness is a morphological feature frequently associated with LAEs, as further discussed in Section 5.3. Given the limited background in these images, the black-level calibration applied during preprocessing may not have been sufficient, potentially affecting the contrast and features the CNN relies on.

TABLE 5.2: Additional characteristics for the catastrophic failures in nLAEs, all features are from COSMOS2020, in order ID, photometric redshift (z_{phot}), I-band magnitude and type of detection

ID	$z_{\textbf{phot}}$	**I-band Magnitude**	**Type**
449920	2.33	24.58	Galaxy
1088508	2.77	25.58	Galaxy
1429856	2.09	25.37	Galaxy
238435	2.30	25.72	Galaxy
646115	2.12	25.78	Galaxy

We extended this analysis across all seven datasets. Table 5.3 presents the top five nLAE sources with the highest prediction error for each dataset. These are additional examples where the model exhibited high confidence despite incorrect predictions. While the photometric redshifts for these sources vary, a considerable number lie in the expected range of LAEs, which may explain their confusion.

TABLE 5.3: catastrophic failures over test set for each dataset. Each row shows source ID from COSMOS2020, dataset, predicted probability, original label, and COSMOS2020 photometric redshift (z_{phot}).

ID	dataset	prob	Original label	z_{phot}
968126	2	0.113	nLAE	2.3336
261457	2	0.269	nLAE	2.0547
1240532	2	0.270	nLAE	2.4804
1032976	2	0.281	nLAE	2.3999
494835	2	0.287	nLAE	2.3392
780300	3	0.266	nLAE	5.4621
1181647	3	0.275	nLAE	2.8136
1282519	3	0.362	nLAE	2.7588
1421483	3	0.058	nLAE	5.2143
477419	3	0.076	nLAE	3.0536
633107	4	0.289	nLAE	5.3799
286967	4	0.079	nLAE	2.1565
381141	4	0.099	nLAE	2.4715
519387	4	0.109	nLAE	5.2909
1369189	4	0.120	nLAE	2.1170
1283357	5	0.293	nLAE	4.3753
909664	5	0.063	nLAE	3.0547
850212	5	0.116	nLAE	2.2765
604174	5	0.126	nLAE	2.4289
442841	5	0.169	nLAE	3.0767
1284646	6	0.240	nLAE	2.1275
443616	6	0.244	nLAE	2.6799
606823	6	0.095	nLAE	2.3393
594269	6	0.181	nLAE	2.4510
262013	6	0.184	nLAE	2.1163
374764	7	0.085	nLAE	2.3891
750374	7	0.125	nLAE	2.1535
713788	7	0.128	nLAE	2.6965
694188	7	0.149	nLAE	2.2149
1278510	7	0.223	nLAE	2.0857

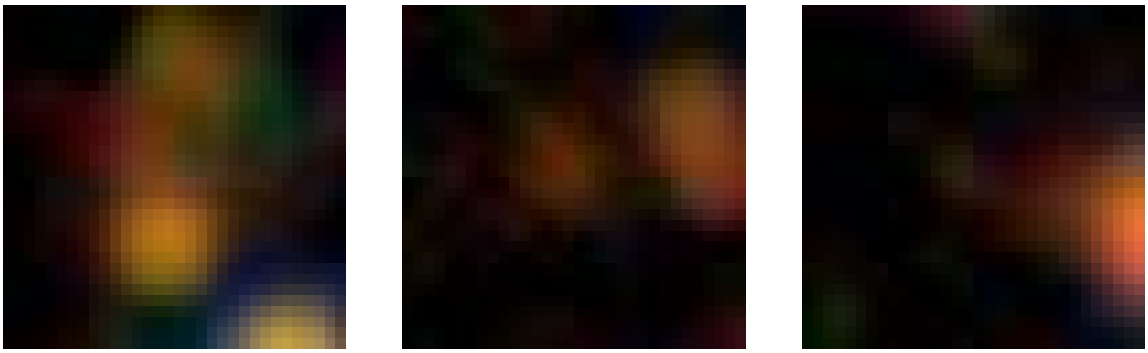
The top 5 most confident misclassifications for LAEs, where true LAEs were predicted as nLAEs, are presented in Table 5.4. These catastrophic failures are characterized by high

predicted probabilities (close to 1.0), which indicate high model confidence in a wrong prediction.

TABLE 5.4: Top 5 catastrophic failures for LAEs, predicted class nLAE and true label LAE, obtained from My_CNN after tune. The table includes the COSMOS2020 ID, the model's predicted probability, the assigned class, and the true class label based on the SC4K.

ID	Predicted probability	Predicted class	True label
326638	0.9646	nLAE (1.0)	LAE (0.0)
1176705	0.9194	nLAE (1.0)	LAE (0.0)
1495768	0.8929	nLAE (1.0)	LAE (0.0)
1220915	0.8919	nLAE (1.0)	LAE (0.0)
511315	0.8764	nLAE (1.0)	LAE (0.0)

Figure 5.2 displays the RGB images for each of these top LAE misclassifications. These visual inspections help assess whether morphological or photometric properties may have led to the model's erroneous decisions.



(A) ID: 326638 (B) ID: 1176705 (C) ID: 1495768

(D) ID: 1220915 (E) ID: 511315

FIGURE 5.2: RGB images of the top catastrophic failures for LAEs. All shown sources are true LAEs that were confidently misclassified as nLAEs.

Upon examining the images, a common trend appears: all five sources seem to be contaminated by nearby or overlapping sources, possibly introducing misleading flux or morphology. Furthermore, as observed in the nLAE analysis, these frames also contain minimal background, which may impair contrast during normalization and feature extraction.

Table 5.5 presents additional astrophysical characteristics of these sources, offering further context for their classification difficulty. Notably, none of the sources have x-ray or radio detections from the SC4K, and most possess strong Lyman-α luminosities and equivalent widths, which usually align with LAE characteristics.

TABLE 5.5: Additional characteristics for the top 5 catastrophic failures among LAEs. Values include ID from COSMOS2020, Lyman-α luminosity ($L_{Ly\alpha}$)(from SC4K), equivalent width (EW_0)(from SC4K), SC4K redshift, X-ray and radio detection flags (from SC4K), and lastly the photometric redshift from COSMOS2020(z_{phot}), some sources from SC4K do not have a photometric redshift, such as ID:1495768.

ID	$\log_{10}$ $(L_{Ly\alpha}/erg\ s^{-1})$	EW_0 $(\AA)$	$z_{\textbf{SC4K}}$	**Radio**	**xRay**	z_{phot}
326638	42.9917	280.85742	3.9	no	no	3.8054
1176705	42.7816	226.30571	3.1	no	no	3.0893
1495768	42.8781	57.660590	5.7	no	no	-
1220915	43.2163	133.32877	4.7	no	no	4.2370
511315	42.9980	90.821320	3.1	no	no	3.0228

As the LAE dataset remained fixed across all seven training iterations, while the nLAE subset was varied, it is expected that the same LAE sources might reappear in multiple experiments. Table 5.6 compiles the top LAE predictions for each dataset and reveals this expected redundancy. Recurrent sources suggest consistent model behavior across independent trainings, potentially pointing to intrinsic ambiguities in their image morphology or photometric context.

For example, source **326638** appears in datasets **2**, **3**, **4**, and **5**, while **760745** also reoccurs in those same datasets. This consistency suggests that the model systematically identifies similar feature patterns in these objects, despite varying negative examples. Similarly, source **173589** is flagged in four different datasets, reinforcing the notion that certain LAEs lie near the decision boundary, becoming highly sensitive to small variations in the training distribution. These results emphasize the need for more diversified labeled LAE samples to reduce overfitting to specific morphologies or observational artifacts.

TABLE 5.6: Top LAE predictions across the seven datasets. Each row contains COSMOS2020 ID, dataset number, prediction probability (prob), Lyman-α luminosity ($L_{Ly\alpha}$) (from SC4K), redshift from SC4K (z_{SC4K}), equivalent width (EW_0) (from SC4K), radio and X-ray detection flags (from SC4K), and the photometric redshift from COSMOS2020 (z_{phot})Repetitions indicate model consistency across different training scenarios.

ID	**dataset**	**prob**	$\log_{10}$ $(L_{Ly\alpha}/erg\ s^{-1})$	$z_{\mathbf{SC4K}}$	EW_0 $(\AA)$	**Radio**	**xRay**	$z_{\mathbf{phot}}$
166452	2	0.856	42.7748	3.40	89.431940	no	no	———
1232790	2	0.836	42.8919	3.01	692.76923	no	no	2.7464
1245590	2	0.834	43.0880	4.08	161.09081	no	no	3.9725
326638	2	0.834	42.8317	3.17	82.852780	no	no	———
760745	2	0.857	43.2800	4.48	411.56403	yes	no	3.9878
1176705	3	0.887	43.0986	4.60	349.75409	no	no	4.4715
326638	3	0.882	42.8317	3.17	82.852780	no	no	———
1232790	3	0.881	42.8919	3.01	692.76923	no	no	2.7464
173589	3	0.880	42.6747	3.40	1015.84485	no	no	———
760745	3	0.880	43.2800	4.48	411.56403	yes	no	3.9878
173589	4	0.850	42.6747	3.40	1015.84485	no	no	———
1220915	4	0.850	43.3779	4.48	81.907880	no	no	———
1232790	4	0.850	42.8919	3.01	692.76923	no	no	2.7464
1491967	4	0.850	43.0263	4.60	117.25982	no	no	———
326638	4	0.850	42.8317	3.17	82.852780	no	no	———
760745	5	0.970	43.2800	4.48	411.56403	yes	no	3.9878
326638	5	0.965	42.8317	3.17	82.852780	no	no	———
1379702	5	0.942	42.7465	3.17	52.309050	no	no	———
1176705	5	0.916	43.0986	4.60	349.75409	no	no	4.4715
405686	5	0.915	43.1485	4.77	157.72897	no	no	0.5270
1402384	6	0.889	42.8381	5.70	512.16528	no	no	0.2544
563030	6	0.888	43.0870	3.74	124.19489	no	no	3.5006
173589	6	0.888	42.6747	3.40	1015.84485	no	no	———
1096673	6	0.888	42.9698	4.08	52.64120	no	no	3.9718
1410092	6	0.888	42.8734	5.70	380.98660	no	no	1.2134
1402384	7	0.949	42.8381	5.70	512.16528	no	no	0.2544
371340	7	0.908	42.8034	3.74	66.662250	no	no	3.1576
173589	7	0.908	42.6747	3.40	1015.8449	no	no	———
217800	7	0.908	43.2588	4.77	706.83423	no	no	0.5950
166452	7	0.904	42.7748	3.40	89.431940	no	no	———

5.2 Impact of data perturbations in the model performance

To assess the robustness of our CNN, I introduced perturbations by training, validating, and testing on varied nLAE datasets with similar but distinct sources. This approach was intended to evaluate the consistency of the model's predictions across different datasets and to understand how slight variations in input data influence model performance.

A total of seven models were generated, resulting in 24,640 sources simultaneously predicted as LAEs across all runs. It is considered as LAE any source in which the prediction value is below 0.5. Table 5.7 displays the distribution of the sources in each bin of probability interval. A total of 45,194 unique LAE prediction after combining the predictions of all models, only marginally different than the 44,295 using only the first dataset model. With a more conservative approach using a threshold of 0.1, the number of unique predictions would have been 915 over the 7 datasets.

TABLE 5.7: Distribution of prediction probabilities across 10 bins of probability values, based on the median score per source over all seven models. This shows the spread of classification confidence and highlights the proportion of high-confidence LAE candidates.

Bin(probability)	Amount
[0.0, 0.1)	915
[0.1, 0.2)	3185
[0.2, 0.3)	6828
[0.3, 0.4)	12598
[0.4, 0.5)	21668
[0.5, 0.6)	26728
[0.6, 0.7)	22957
[0.7, 0.8)	25928
[0.8, 0.9)	65760
[0.9, 1.0)	5259

The metrics were similar across the datasets, which shows that even after training different models with varying nLAEs sources, the overall performance was maintained. The average accuracy was 75.84%, as shown in Figure 5.3, and the average F1-score was 75.51%, as shown in Figure 5.4.

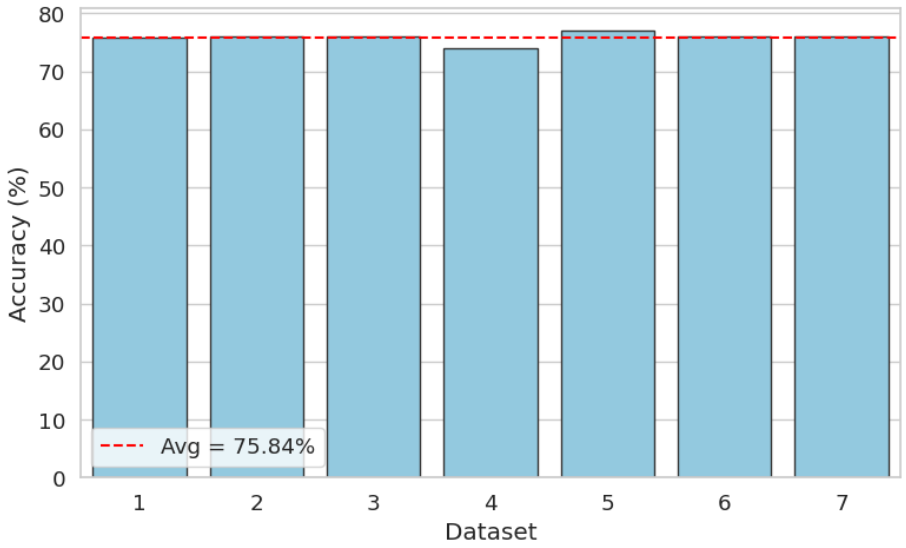


FIGURE 5.3: Accuracy across all datasets, showing that despite varying nLAEs sources the metrics overall are the same.

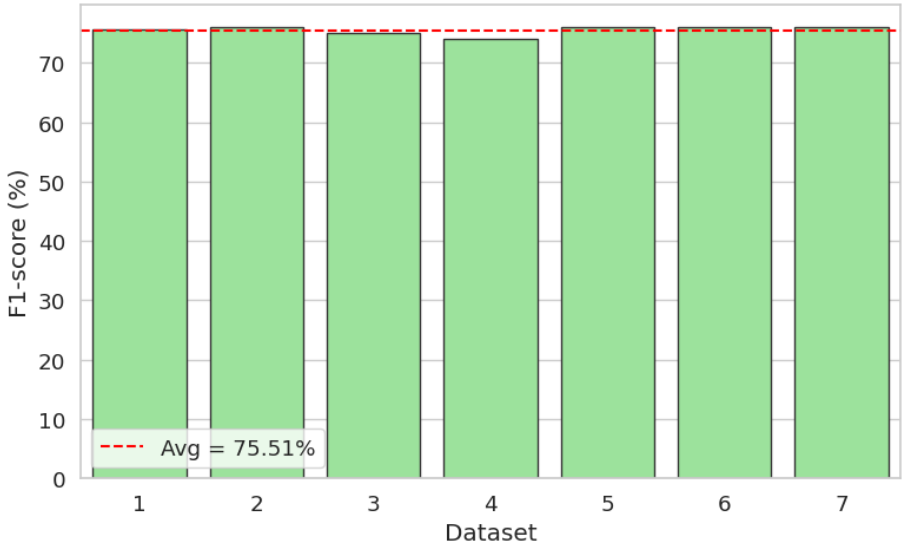
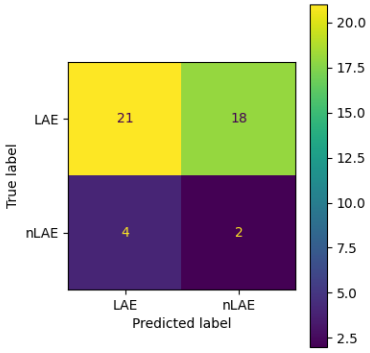
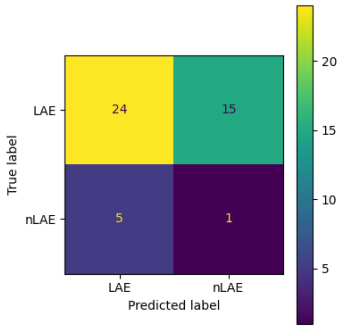


FIGURE 5.4: F1-score across all datasets. Consistent with the accuracy, all datasets showed similar F1-score values.

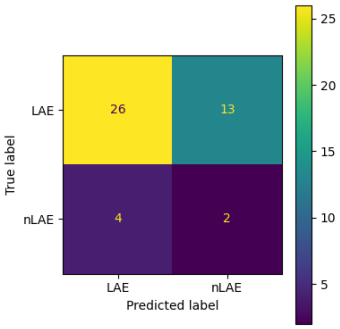
To further evaluate the performance of the perturbed CNN, the predictions were cross-matched with the HETDEX [109] survey, in a similar approach as in the classification results, which provides spectroscopic confirmation of LAEs. Out of the 45 matched sources, the confusion matrices for each perturbed dataset are shown in Figures 5.5a to 5.5g. The performance varied across the perturbed datasets, with true positive rates ranging from 21 to 26 and false negative rates from 13 to 18.



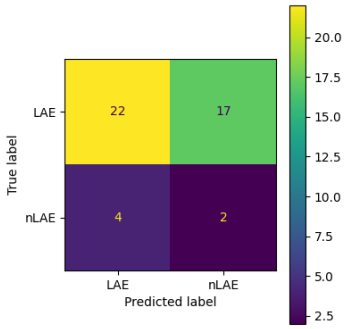
(A) Dataset 1.



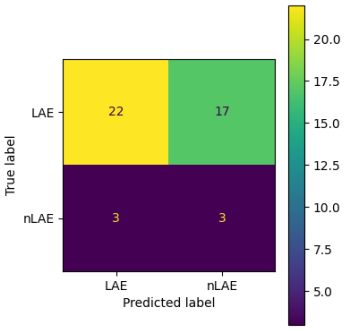
(B) Dataset 2.



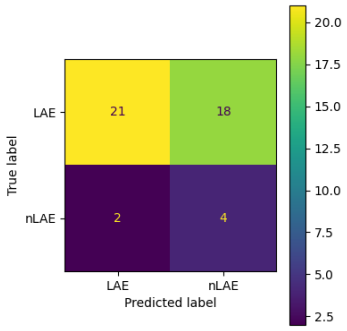
(C) Dataset 3.



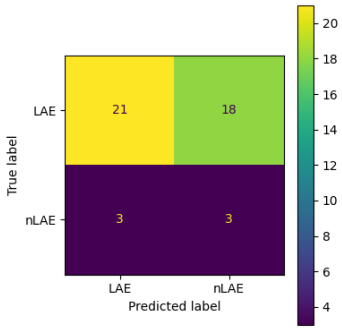
(D) Dataset 4.



(E) Dataset 5.



(F) Dataset 6.



(G) Dataset 7.

FIGURE 5.5: Confusion matrices for each of the seven classification datasets, evaluated using 45 crossmatched sources from the HETDEX survey. These visualizations highlight the consistency and variation in classification outcomes across different negative (nLAE) sample configurations.

The correlation matrix of predictions across them (Figure 5.6) shows high correlation among the predictions from different datasets, with values ranging from 0.85 to 0.95. This underscores the insensitivity of the models to the perturbations applied.

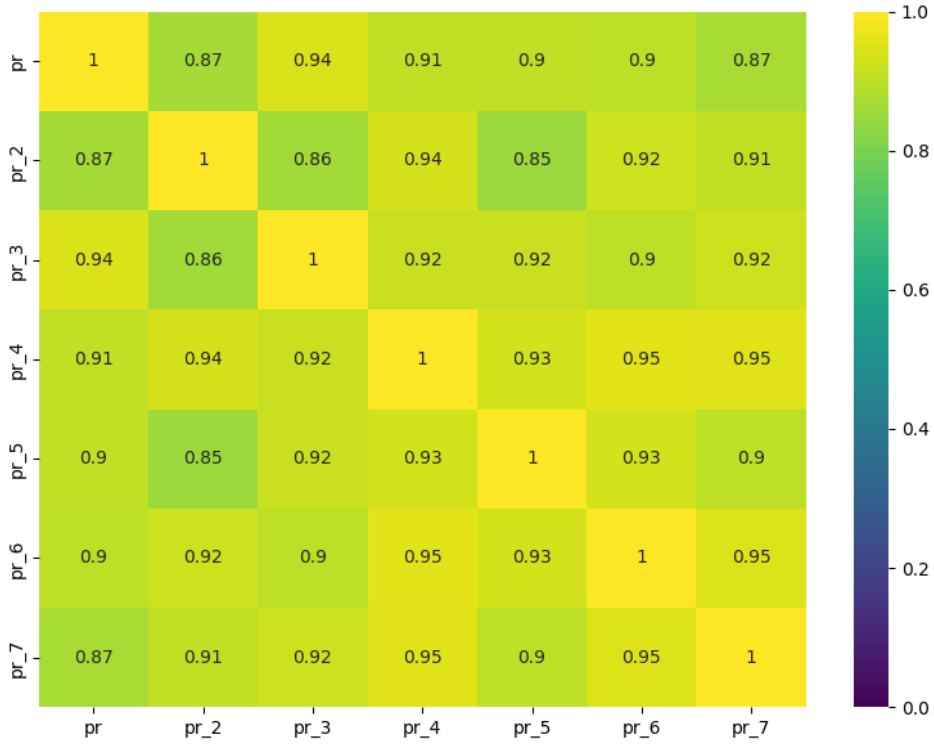


FIGURE 5.6: Correlation matrix of classification probabilities across the seven perturbed datasets. Each cell indicates the Pearson correlation coefficient between the predicted probabilities from different models, revealing strong consistency despite the changes in negative class composition.

Table 5.8 compares key metrics among all datasets. Notably, precision remained high across all perturbed datasets, indicating that it is usually correct when the model predicts a source as LAE. However, recall had lower values with range from 53% to 66% while accuracy reaching a maximum of 62%, and F1-score ranging in 65% to 75%. These results show that despite changing the nLAE dataset each models achieve similar results.

TABLE 5.8: Summary of classification performance metrics for the original dataset (Dataset 1) and each of the six perturbed datasets over the HETDEX crossmatched sources. The table reports true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), along with derived metrics: precision, recall, accuracy, and F1-score (all in percentage). Despite variations in the nLAE class across datasets, the models maintain high precision and comparable F1-scores, though recall and overall accuracy remain limited due to false negative rates.

Dataset	TP	TN	FP	FN	Precision (%)	Recall (%)	Accuracy (%)	F1-Score (%)
1	21	2	4	18	84.00	53.85	51.11	65.62
2	24	1	5	15	82.76	61.54	55.56	70.59
3	26	2	4	13	86.67	66.67	62.22	75.36
4	22	2	4	17	84.62	56.41	53.33	67.69
5	22	3	3	17	88.00	56.41	55.56	68.75
6	21	4	2	18	91.30	53.85	55.56	67.74
7	21	3	3	18	87.50	53.85	53.33	66.67
Mean	–	–	–	–	**86.41**	**57.51**	**55.24**	**68.92**

- **Precision consistency:** Across all perturbed datasets, precision remains high, with a maximum of 91.30%. This indicates that the model's positive predictions are generally accurate, even when trained on different datasets with similar data.
- **Recall:** All datasets had lower values for recall compared with precision, which displays the difficulty in classifying a source as nLAE.
- **Accuracy:** The highest accuracy observed is 62.22% which is in general a low value, and that is caused by the bad nLAE predicitons metrics.
- **F1-Score:** The F1-Score, balancing precision and recall, varies across the datasets, reaching a maximum of 75.36% and a minimum of 65.62%.

The analysis of perturbation effects on model predictions reveals the significant impact of training data variability on the stability and reliability of LAE classification. The similarities observed across different datasets underscore the models' low sensitivity to changes in the input data, more specifically the nLAE data.

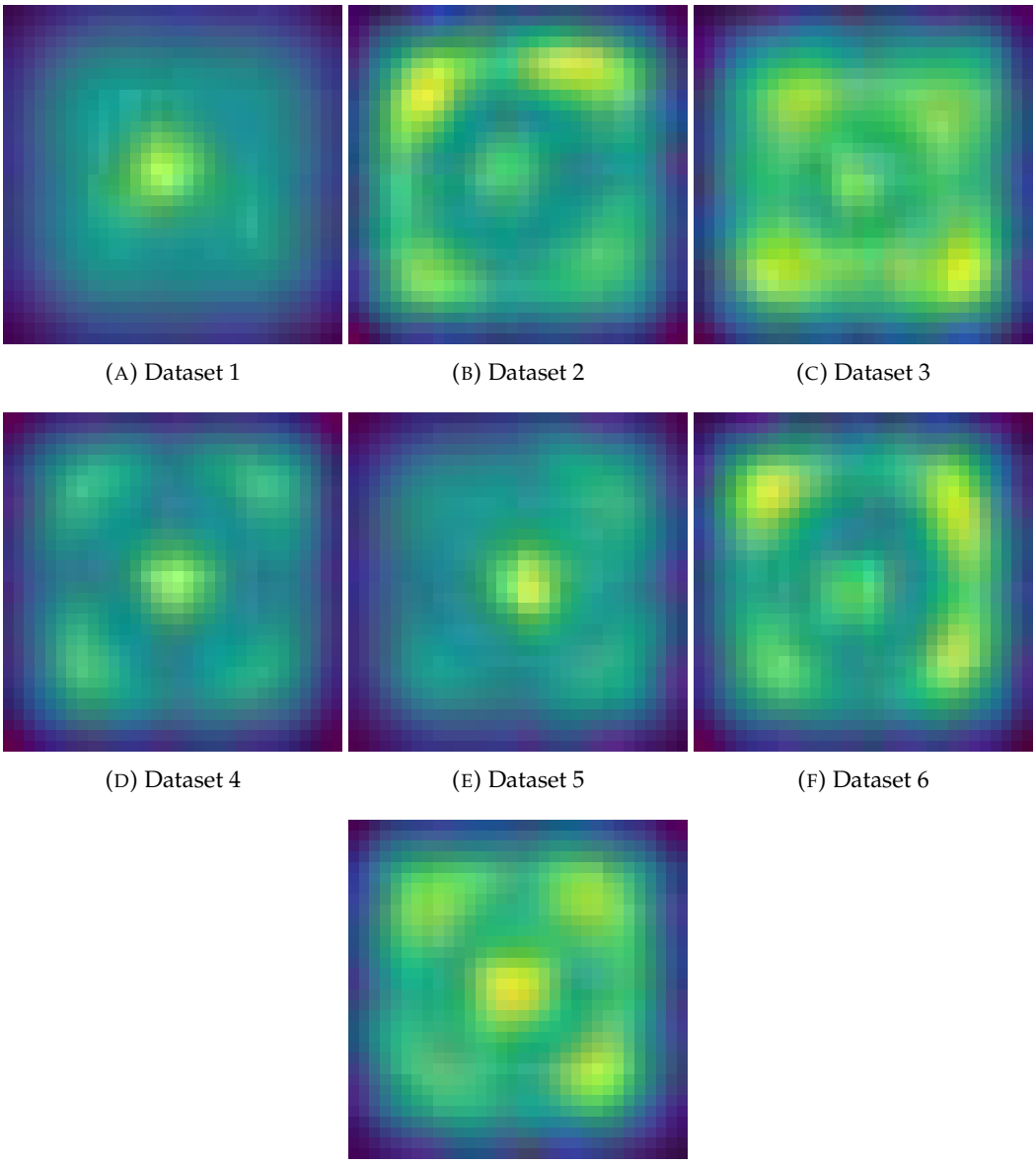
Furthermore, the cross-match results with the HETDEX [109] survey provide a benchmark for evaluating the models' real-world applicability. Although some perturbed datasets showed improvements in certain metrics, the overall variability suggests that, despite being precise, the CNN misses some LAEs, and is overall conservative, which could be improved with more LAE sources to train the architecture, or more information encoded as color from other filters, or feature values.

5.3 Interpretability analysis of CNN activations

This section presents the interpretability analysis of the CNN used for classification. By leveraging saliency maps, this work aims to uncover the underlying mechanisms by which the CNN makes its predictions. Saliency maps highlight the regions of the input image that most strongly influence the model's output by computing the gradient of the prediction with respect to the input pixels. Understanding these mechanisms is essential, as they may correspond to physical conditions that facilitate the escape of Lyman-α photons

Figure 5.7 shows the stacked saliency maps generated from the LAE test set for each dataset. Although the same LAE sources are used across all datasets, each model interprets them slightly differently. However, a common pattern emerges: most maps exhibit strong activation centered on the main source, along with some surrounding activation. This suggests that features related to the central galaxy and its spatial compactness are the primary focus of the model.

Figure 5.8 shows the equivalent maps for nLAEs. Although the sources vary across datasets, a similar structure appears. These maps tend to show little or no central activation, with most saliency concentrated in the periphery. This suggests the model bases its classification of nLAEs more on surrounding features or background patterns rather than properties of a central object.

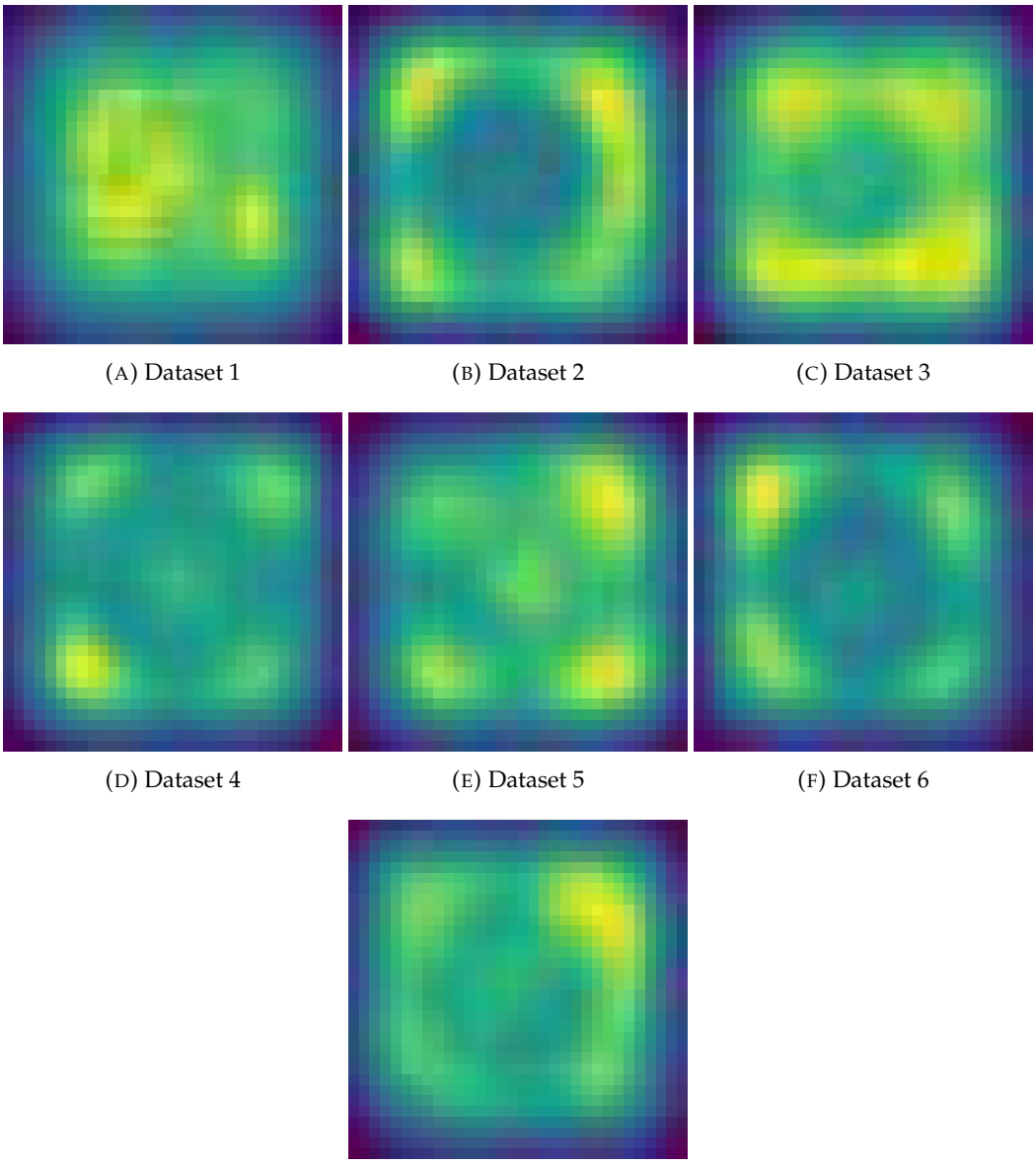


(A) Dataset 1 (B) Dataset 2 (C) Dataset 3

(D) Dataset 4 (E) Dataset 5 (F) Dataset 6

(G) Dataset 7

FIGURE 5.7: Stacked saliency maps for LAE sources across all seven datasets. Most show central activation, highlighting the importance of the core region and compact morphology in the CNN's decision process.



(A) Dataset 1 (B) Dataset 2 (C) Dataset 3

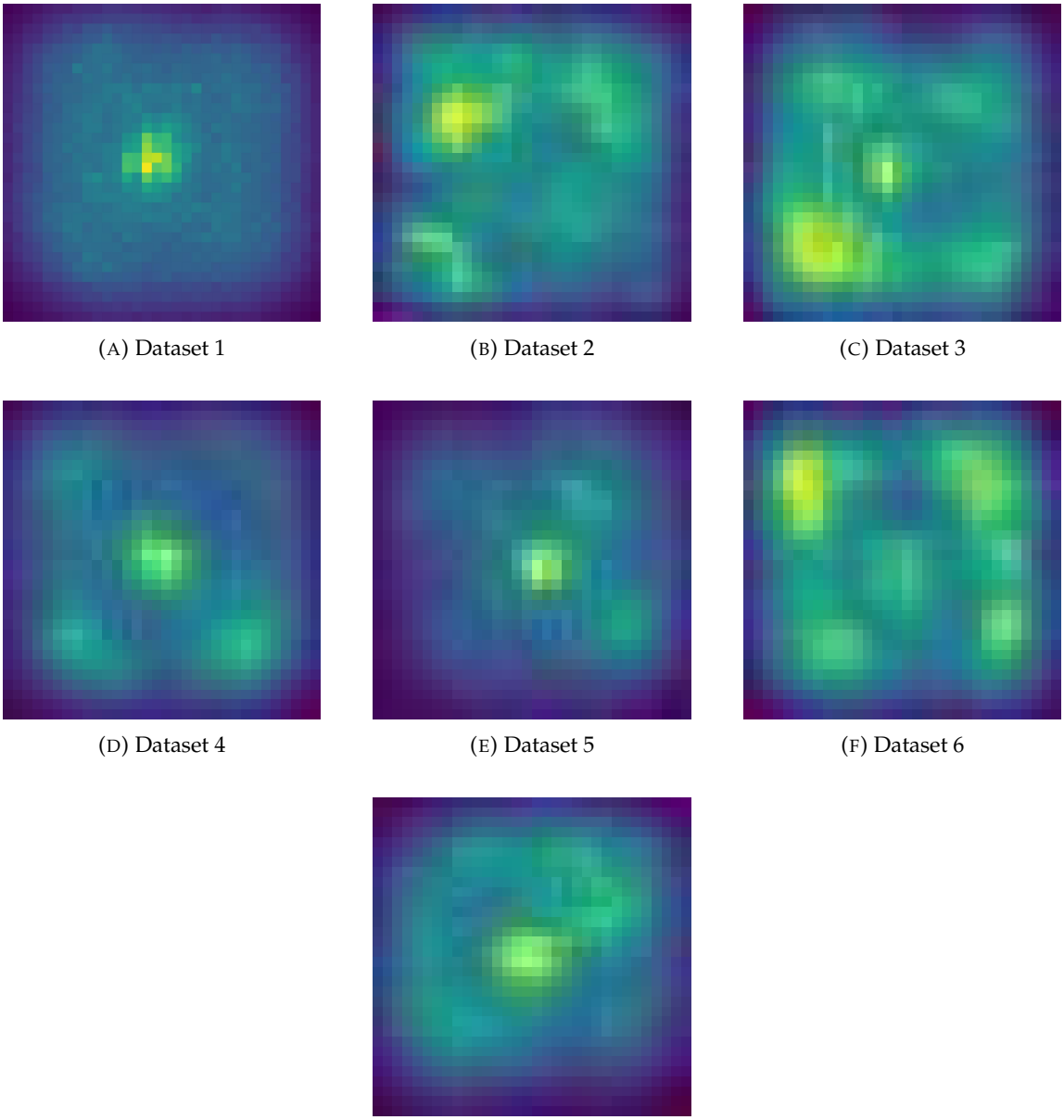
(D) Dataset 4 (E) Dataset 5 (F) Dataset 6

(G) Dataset 7

FIGURE 5.8: Stacked saliency maps for nLAE sources across all datasets. Activations are typically off-center, suggesting that classification is based on the presence or absence of peripheral or background features rather than a prominent central source.

From these saliency map stacks, one can infer that the CNN primarily relies on spatial structure for its classification decisions. For nLAEs, the model appears to react to missing or diffuse information around the central area, while for LAEs, central compactness and brightness seem to be the key signals.

To further explore the classification failures, we examine stacked saliency maps of the top five misclassified sources from each dataset. Figure 5.9 shows the stack of catastrophic nLAE errors that were incorrectly predicted as LAEs. These errors typically exhibit central activations, contrary to most nLAEs, suggesting that the CNN was ”tricked” by sources mimicking LAE-like central brightness.



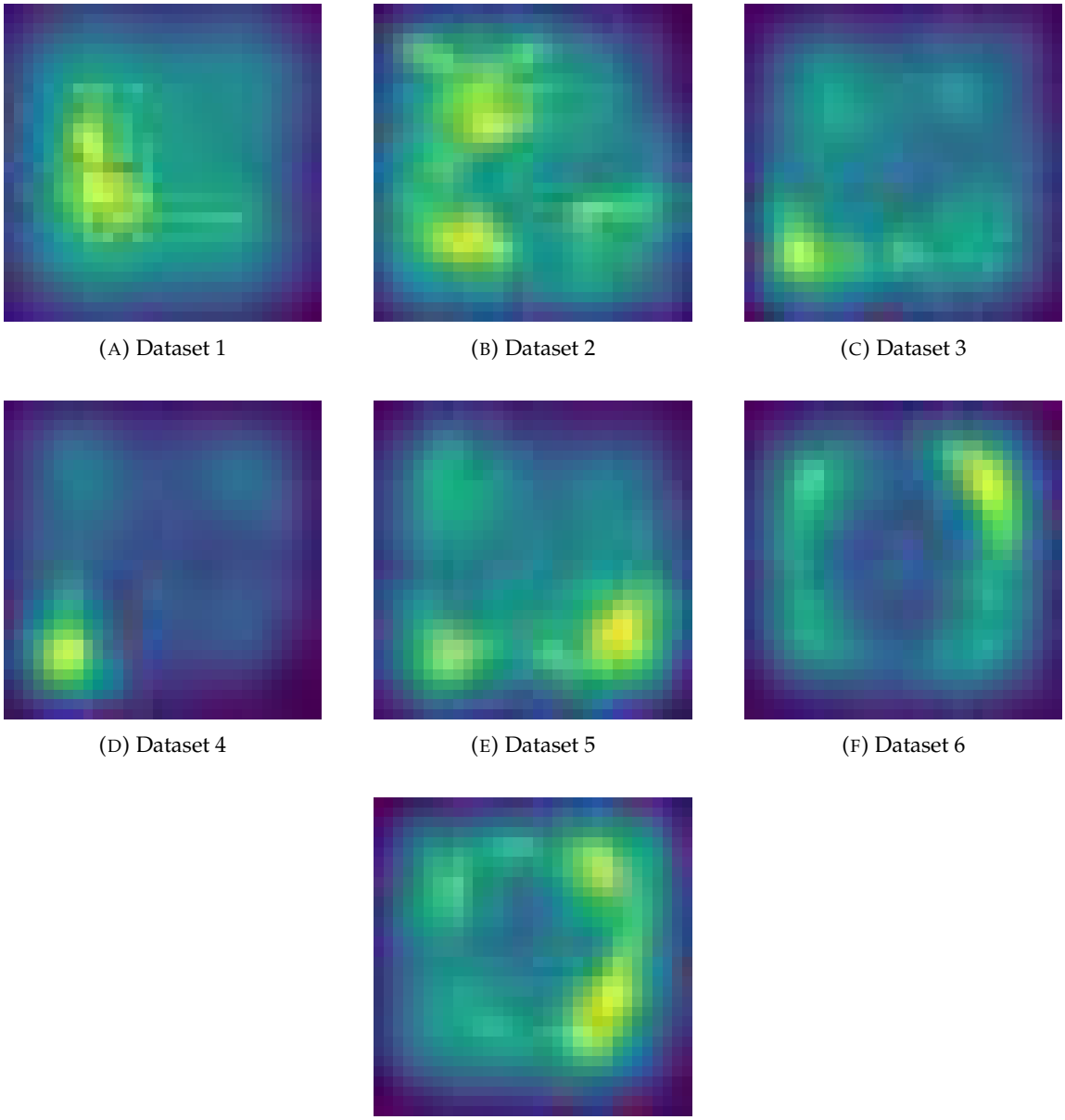
(A) Dataset 1 (B) Dataset 2 (C) Dataset 3

(D) Dataset 4 (E) Dataset 5 (F) Dataset 6

(G) Dataset 7

FIGURE 5.9: Stacked saliency maps of the top catastrophic nLAE misclassifications across all datasets. Although the true label is nLAE, the CNN predicted LAE, likely due to the presence of central activation together with some peripheral activation resembling compact LAEs.

Figure 5.10 presents the corresponding maps for LAEs misclassified as nLAEs. These typically show weak or no activation in the center and stronger signals in the surrounding area. This may reflect confusion caused by nearby sources, making the central LAE less distinct to the CNN.



(A) Dataset 1 (B) Dataset 2 (C) Dataset 3

(D) Dataset 4 (E) Dataset 5 (F) Dataset 6

(G) Dataset 7

FIGURE 5.10: Stacked saliency maps of the top catastrophic LAE misclassifications across all datasets. Although the true label is LAE, the CNN predicted nLAE, often due to a lack of central activation, possibly caused by blending or confusion with nearby sources.

Chapter 6

Conclusion and future directions

This work developed and tested Deep Learning models to identify and extract physical properties of Lyman-α emitters, using only broadband photometric data. Through the application of CNNs to RGB images constructed from g, r, and i bands, it was possible to classify sources as LAEs or nLAEs, and predict features such as redshift, Lyman-α luminosity, and equivalent width.

The classification model achieved an average accuracy of 75.84%, demonstrating the feasibility of detecting LAEs similar to those in the SC4K sample using limited photometric information. The regression models predicted redshift and Lyman-α luminosity (with robust scaling) and equivalent width (with log scaling) with competitive performance, achieving average MAE as low as 0.032. These results highlight the importance of preprocessing strategies and input representation in optimizing model outcomes.

A comparison between multiple known CNN architectures and custom models showed that task-specific designs can outperform general-purpose deep networks when data is limited in spectral depth or resolution. Despite working with just three photometric bands, the CNNs effectively learned spatial patterns that distinguish LAEs from nLAEs.

The perturbation analysis, conducted by training seven CNNs with identical architectures but varied nLAE datasets, revealed the robustness of the model predictions. A total of 24,640 sources were consistently classified as LAEs across all models, and performance metrics such as F1-score remained stable, with an average value of 75.51% over the test dataset. This confirms that the models are largely insensitive to small changes in the negative class sampling.

Analysis of saliency maps revealed that the models focus strongly on the central region

of LAEs, usually compact and bright, while nLAEs show more diffuse or peripheral activation. This spatial pattern is consistent with physical expectations and literature reports that LAEs tend to be more compact than typical nLAEs [110]. Catastrophic classification errors further support this conclusion: most occurred when an LAE had saliency maps resembling nLAEs (i.e., lacking central activation) or vice versa.

These findings reinforce the conclusion that the CNNs are not merely "black boxes" but instead learn physically interpretable features, such as spatial size and central brightness. Nevertheless, misclassifications suggest that some information necessary for perfect separation may be missing from the limited gri bands or lost during preprocessing.

Although the saliency analysis was only applied to the classification models, extending this approach to the regression networks could yield additional insight into how the CNN estimates continuous physical parameters and where these predictions may fail.

Future directions: To enhance the classification and regression capabilities, future efforts could focus on several aspects. First, integrating additional photometric bands or constructing composite features that preserve RGB format could enrich the input information. Second, improving background calibration (e.g., black-level subtraction) and exploring different image resolutions may help capture finer features. Third, applying and comparing multiple saliency methods, could further elucidate the model's decision-making. Finally, the generalizability of these models to new surveys, such as Euclid [31], should be tested. These CNNs may serve as a foundation for scalable candidate selection and first-guess characterization in large photometric datasets.

In conclusion, this study demonstrates the potential of deep learning models to extract meaningful astrophysical information from minimal data. With careful model design and interpretation, CNNs can become powerful tools in the search and study of high-redshift galaxies in current and upcoming surveys.

Bibliography

[1] D. Sobral, S. Santos, J. Matthee, A. Paulino-Afonso, B. Ribeiro, J. Calhau, and A. A. Khostovan, "Slicing COSMOS with SC4K: The Evolution of Typical Lyα Emitters and the Lyα Escape Fraction from $z \sim 2$ to 6 ," *Monthly Notices of the Royal Astronomical Society*, vol. 476, no. 4, pp. 4725–4752, feb 2018. [Online]. Available: http://dx.doi.org/10.1093/mnras/sty378 [Cited on pages ii, xi, 4, 13, 15, 16, 17, and 51.]

[2] B. Draine, *Physics of the Interstellar and Intergalactic Medium* , ser. Princeton Series in Astrophysics. Princeton University Press, 2010. [Online]. Available: https://books.google.pt/books?id=FycJvKHyiwsC [Cited on page 1.]

[3] M. Hayes, "Lyman Alpha Emitting Galaxies in the Nearby Universe," *Publications of the Astronomical Society of Australia*, vol. 32, 2015. [Online]. Available: http://dx.doi.org/10.1017/pasa.2015.25 [Cited on pages 1 and 2.]

[4] R. B. Partridge and P. J. E. Peebles, "Are Young Galaxies Visible? ," , vol. 147, p. 868, mar 1967. [Cited on pages 1 and 2.]

[5] K. E. Heintz, D. Watson, G. Brammer, S. Vejlgaard, A. Hutter, V. B. Strait, J. Matthee, P. A. Oesch, P. Jakobsson, N. R. Tanvir, P. Laursen, R. P. Naidu, C. A. Mason, M. Killi, I. Jung, T. Y.-Y. Hsiao, Abdurro'uf, D. Coe, P. A. Haro, S. L. Finkelstein, and S. Toft, "Extreme damped Lyman-α absorption in young star-forming galaxies at z = 9-11 ," 2023. [Online]. Available: https://arxiv.org/abs/2306.00647 [Cited on page 1.]

[6] K. K. Nilsson, "The Lyman-alpha Emission Line as a Cosmological Tool," Ph.D. dissertation, Niels Bohr Institute for Astronomy, Physics and Geophysics, nov 2007. [Cited on page 1.]

[7] Y. Huang, K.-S. Lee, O. Cucciati, B. C. Lemaux, M. Sawicki, N. Malavasi, V. Ramakrishnan, R. Xue, L. P. Cassara, Y.-K. Chiang, A. Dey, S. D. J. Gwyn, N. Hathi, L. Pentericci, M. K. M. Prescott, and G. Zamorani, "Evaluating Lyα Emission as a Tracer of the Largest Cosmic Structure at $z - 2.47$," , vol. 941, no. 2, p. 134, dec 2022. [Cited on page 1.]

[8] M. Ouchi, Y. Ono, and T. Shibuya, "Observations of the Lyman-α Universe ," *Annual Review of Astronomy and Astrophysics*, vol. 58, no. 1, p. 617–659, aug 2020. [Online]. Available: http://dx.doi.org/10.1146/annurev-astro-032620-021859 [Cited on pages 1, 2, 3, 5, and 6.]

[9] O. Nebrin, A. Smith, K. Lorinc, J. Hörnquist, A. Larson, G. Mellema, and S. K. Giri, "Lyman-α feedback prevails at Cosmic Dawn: implications for the first galaxies, stars, and star clusters ," *Monthly Notices of the Royal Astronomical Society*, vol. 537, no. 2, p. 1646–1687, jan 2025. [Online]. Available: http://dx.doi.org/10.1093/mnras/staf038 [Cited on page 2.]

[10] G. A. Oyarzún, G. A. Blanc, V. González, M. Mateo, J. I. B. III, S. L. Finkelstein, P. Lira, J. D. Crane, and E. W. Olszewski, "How Lymanα Emission Depends on Galaxy Stellar Mass ," *The Astrophysical Journal Letters*, vol. 821, no. 1, p. L14, apr 2016. [Online]. Available: https://dx.doi.org/10.3847/2041-8205/821/1/L14 [Cited on page 2.]

[11] M. Mori and M. Umemura, "Evolution of Lyman-α Emitters, Lyman-break Galaxies and Elliptical Galaxies ," in *Panoramic Views of Galaxy Formation and Evolution*, ser. Astronomical Society of the Pacific Conference Series, T. Kodama, T. Yamada, and K. Aoki, Eds., vol. 399, oct 2008, p. 288. [Cited on page 2.]

[12] B. Villasenor, B. Robertson, P. Madau, and E. Schneider, "Inferring the Thermal History of the Intergalactic Medium from the Properties of the Hydrogen and Helium Lyα Forest ," *The Astrophysical Journal*, vol. 933, no. 1, p. 59, jul 2022. [Online]. Available: http://dx.doi.org/10.3847/1538-4357/ac704e [Cited on page 2.]

[13] F. Nasir, P. Gaikwad, F. B. Davies, J. S. Bolton, E. Puchwein, and S. E. I. Bosman, "Deep Learning the Intergalactic Medium using Lyman-alpha Forest at $4 \leq z \leq 5$," 2024. [Online]. Available: https://arxiv.org/abs/2404.05794 [Cited on page 2.]

[14] M. Dijkstra, "Lyα Emitting Galaxies as a Probe of Reionisation ," *Publications of the Astronomical Society of Australia*, vol. 31, 2014. [Online]. Available: http://dx.doi.org/10.1017/pasa.2014.33 [Cited on pages 2, 5, and 6.]

[15] C. Witten, N. Laporte, S. Martin-Alvarez, D. Sijacki, Y. Yuan, M. G. Haehnelt, W. M. Baker, J. S. Dunlop, R. S. Ellis, N. A. Grogin, G. Illingworth, H. Katz, A. M. Koekemoer, D. Magee, R. Maiolino, W. McClymont, P. G. Pérez-González, D. Puskás, G. Roberts-Borsani, P. Santini, and C. Simmonds, "Deciphering Lyman-α emission deep into the epoch of reionization ," *Nature Astronomy*, vol. 8, no. 3, p. 384–396, jan 2024. [Online]. Available: http://dx.doi.org/10.1038/s41550-023-02179-3 [Cited on page 2.]

[16] Gatuzz, E., Wilms, J., Hämmerich, S., and Arcodia, R., "Probing the physical properties of the intergalactic medium using SRG/eROSITA spectra from blazars ," *AA*, vol. 683, p. A213, 2024. [Online]. Available: https://doi.org/10.1051/0004-6361/202348705 [Cited on page 2.]

[17] A. A. Meiksin, "The physics of the intergalactic medium ," *Reviews of Modern Physics*, vol. 81, no. 4, p. 1405–1469, oct 2009. [Online]. Available: http://dx.doi.org/10.1103/RevModPhys.81.1405 [Cited on page 2.]

[18] E. M. Hu and R. G. McMahon, "Detection of Lyman-α-emitting galaxies at redshift 4.55 ," *Nature*, vol. 382, no. 6588, p. 231–233, jul 1996. [Online]. Available: http://dx.doi.org/10.1038/382231a0 [Cited on page 2.]

[19] S. C. Odewahn, R. A. Windhorst, S. P. Driver, and W. C. Keel, "Automated Morphological Classification in Deep Hubble Space Telescope UBVI Fields: Rapidly and Passively Evolving Faint Galaxy Populations ," *The Astrophysical Journal*, vol. 472, no. 1, p. L13, nov 1996. [Online]. Available: https://dx.doi.org/10.1086/310345 [Cited on page 2.]

[20] L. L. Cowie and E. M. Hu, "High-z Lyα Emitters. I. A Blank-Field Search for Objects near Redshift z = 3.4 in and around the Hubble Deep Field and the Hawaii Deep Field SSA 22 ," *The Astronomical Journal*, vol. 115, no. 4, p. 1319–1328, apr 1998. [Online]. Available: http://dx.doi.org/10.1086/300309 [Cited on page 2.]

[21] I. S. McLean, C. C. Steidel, H. W. Epps, N. Konidaris, K. Y. Matthews, S. Adkins, T. Aliado, G. Brims, J. M. Canfield, J. L. Cromer, J. Fucik, K. Kulas, G. Mace,

K. Magnone, H. Rodriguez, G. Rudie, R. Trainor, E. Wang, B. Weber, and J. Weiss, "MOSFIRE, the multi-object spectrometer for infra-red exploration at the Keck Observatory ," in *Ground-based and Airborne Instrumentation for Astronomy IV*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, I. S. McLean, S. K. Ramsay, and H. Takami, Eds., vol. 8446, sep 2012, p. 84460J. [Cited on page 3.]

[22] D. C. Martin, J. Fanson, D. Schiminovich, P. Morrissey, P. G. Friedman, T. A. Barlow, T. Conrow, R. Grange, P. N. Jelinsky, B. Milliard, O. H. W. Siegmund, L. Bianchi, Y.-I. Byun, J. Donas, K. Forster, T. M. Heckman, Y.-W. Lee, B. F. Madore, R. F. Malina, S. G. Neff, R. M. Rich, T. Small, F. Surber, A. S. Szalay, B. Welsh, and T. K. Wyder, "The Galaxy Evolution Explorer: A Space Ultraviolet Survey Mission ," *The Astrophysical Journal*, vol. 619, no. 1, p. L1–L6, jan 2005. [Online]. Available: http://dx.doi.org/10.1086/426387 [Cited on page 3.]

[23] G. Meylan, J. P. Madrid, and D. Macchetto, "Hubble Space TelescopeScience Metrics ," *Publications of the Astronomical Society of the Pacific*, vol. 116, no. 822, p. 790–796, aug 2004. [Online]. Available: http://dx.doi.org/10.1086/423227 [Cited on pages 3 and 4.]

[24] De Cia, A., Ledoux, C., Mattsson, L., Petitjean, P., Srianand, R., Gavignaud, I., and Jenkins, E. B., "Dust-depletion sequences in damped Lyman-α absorbers - A unified picture from low-metallicity systems to the Galaxy ," *AA*, vol. 596, p. A97, 2016. [Online]. Available: https://doi.org/10.1051/0004-6361/201527895 [Cited on page 3.]

[25] T. Su, "Dusty Star-forming Galaxies at High Redshift ," Ph.D. dissertation, Johns Hopkins University, Maryland, feb 2017. [Cited on page 3.]

[26] M. R. Blanton and J. Moustakas, "Physical Properties and Environments of Nearby Galaxies ," , vol. 47, no. 1, pp. 159–210, sep 2009. [Cited on page 3.]

[27] S. C. Ellis, J. Bland-Hawthorn, J. S. Lawrence, A. J. Horton, R. Content, M. M. Roth, N. Pai, R. Zhelem, S. Case, E. Hernandez, S. G. Leon-Saval, R. Haynes, S. S. Min, D. Giannone, K. Madhav, A. Rahman, C. Betters, D. Haynes, W. Couch, L. J. Kewley, R. McDermid, L. Spitler, R. G. Sharp, and S. Veilleux, "First demonstration of OH suppression in a high-efficiency near-infrared spectrograph ," *Monthly*

Notices of the Royal Astronomical Society, vol. 492, no. 2, pp. 2796–2806, jan 2020. [Online]. Available: https://doi.org/10.1093/mnras/staa028 [Cited on page 3.]

[28] M. Ouchi, *Observations of Ly α Emitters at High Redshift* . Springer Berlin Heidelberg, 2019, p. 189–318. [Online]. Available: http://dx.doi.org/10.1007/978-3-662-59623-4_3 [Cited on pages 4 and 5.]

[29] C. L. Martin, M. Sawicki, A. Dressler, and P. McCarthy, "A Magellan IMACS Spectroscopic Search for Lyα-emitting Galaxies at Redshift 5.7," *The Astrophysical Journal*, vol. 679, no. 2, p. 942–961, jun 2008. [Online]. Available: http://dx.doi.org/10.1086/586729 [Cited on page 4.]

[30] M. Sawicki, B. C. Lemaux, P. Guhathakurta, E. N. Kirby, N. P. Konidaris, C. L. Martin, M. C. Cooper, D. C. Koo, J. A. Newman, and B. J. Weiner, "The DEEP2 Redshift Survey: Lyα Emitters in the Spectroscopic Database ," *The Astrophysical Journal*, vol. 687, no. 2, p. 884–898, nov 2008. [Online]. Available: http://dx.doi.org/10.1086/591779 [Cited on page 4.]

[31] G. D. Racca, R. Laureijs, L. Stagnaro, J.-C. Salvignol, J. Lorenzo Alvarez, G. Saavedra Criado, L. Gaspar Venancio, A. Short, P. Strada, T. Bönke, C. Colombo, A. Calvi, E. Maiorano, O. Piersanti, S. Prezelus, P. Rosato, J. Pinel, H. Rozemeijer, V. Lesna, P. Musi, M. Sias, A. Anselmi, V. Cazaubiel, L. Vaillon, Y. Mellier, J. Amiaux, M. Berthé, M. Sauvage, R. Azzollini, M. Cropper, S. Pottinger, K. Jahnke, A. Ealet, T. Maciaszek, F. Pasian, A. Zacchei, R. Scaramella, J. Hoar, R. Kohley, R. Vavrek, A. Rudolph, and M. Schmidt, "The Euclid mission design ," in *Space Telescopes and Instrumentation 2016: Optical, Infrared, and Millimeter Wave*, H. A. MacEwen, G. G. Fazio, M. Lystrup, N. Batalha, N. Siegler, and E. C. Tong, Eds., vol. 9904. SPIE, jul 2016, p. 99040O. [Online]. Available: http://dx.doi.org/10.1117/12.2230762 [Cited on pages 4, 8, and 84.]

[32] J. P. Gardner, J. C. Mather, M. Clampin, R. Doyon, M. A. Greenhouse, H. B. Hammel, J. B. Hutchings, P. Jakobsen, S. J. Lilly, K. S. Long, J. I. Lunine, M. J. Mccaughrean, M. Mountain, J. Nella, G. H. Rieke, M. J. Rieke, H.-W. Rix, E. P. Smith, G. Sonneborn, M. Stiavelli, H. S. Stockman, R. A. Windhorst, and G. S. Wright, "The James Webb Space Telescope ," *Space Science Reviews*, vol. 123, no. 4, p. 485–606, apr 2006. [Online]. Available: http://dx.doi.org/10.1007/s11214-006-8315-7 [Cited on pages 4 and 8.]

[33] Y. Ning, Z. Cai, X. Lin, Z.-Y. Zheng, X. Feng, M. Li, Q. Li, D. Spinoso, Y. Wu, and H. Zhang, "Unveiling Luminous Lyα Emitters at $z \approx 6$ through JWST/NIRCam Imaging in the COSMOS Field ," 2024. [Online]. Available: https://arxiv.org/abs/2312.04841 [Cited on page 5.]

[34] J. Kerutt, L. Wisotzki, A. Verhamme, K. B. Schmidt, F. Leclercq, E. C. Herenz, T. Urrutia, T. Garel, T. Hashimoto, M. Maseda, J. Matthee, H. Kusakabe, J. Schaye, J. Richard, B. Guiderdoni, V. Mauerhofer, T. Nanayakkara, and E. Vitte, "Equivalent widths of Lyman α emitters in MUSE-Wide and MUSE-Deep ," *Astronomy amp; Astrophysics*, vol. 659, p. A183, mar 2022. [Online]. Available: http://dx.doi.org/10.1051/0004-6361/202141900 [Cited on page 5.]

[35] J. D. Kurk, A. Cimatti, S. di Serego Alighieri, J. Vernet, E. Daddi, A. Ferrara, and B. Ciardi, "A Lymanα emitter at $z = 6.5$ found with slitless spectroscopy ," *Astronomy amp; Astrophysics*, vol. 422, no. 1, p. L13–L17, jul 2004. [Online]. Available: http://dx.doi.org/10.1051/0004-6361:20040189 [Cited on page 5.]

[36] A. Saxena, A. J. Bunker, G. C. Jones, D. P. Stark, A. J. Cameron, J. Witstok, S. Arribas, W. M. Baker, S. Baum, R. Bhatawdekar, R. Bowler, K. Boyett, S. Carniani, S. Charlot, J. Chevallard, M. Curti, E. Curtis-Lake, D. J. Eisenstein, R. Endsley, K. Hainline, J. M. Helton, B. D. Johnson, N. Kumari, T. J. Looser, R. Maiolino, M. Rieke, H.-W. Rix, B. E. Robertson, L. Sandles, C. Simmonds, R. Smit, S. Tacchella, C. C. Williams, C. N. A. Willmer, and C. Willott, "JADES: The production and escape of ionizing photons from faint Lyman-alpha emitters in the epoch of reionization ," 2024. [Online]. Available: https://arxiv.org/abs/2306.04536 [Cited on page 5.]

[37] Raiter, A., Schaerer, D., and Fosbury, R. A. E., "Predicted uv properties of very metal-poor starburst galaxies," *AA*, vol. 523, p. A64, 2010. [Online]. Available: https://doi.org/10.1051/0004-6361/201015236 [Cited on page 5.]

[38] M. A. Latif, D. R. G. Schleicher, M. Spaans, and S. Zaroubi, "Lymanα emission from the first galaxies: signatures of accretion and infall in the presence of line trapping ," , vol. 413, no. 1, pp. L33–L37, may 2011. [Cited on page 6.]

[39] F. van de Voort, J. Schaye, G. Altay, and T. Theuns, "Cold accretion flows and the nature of high column density Hi absorption at redshift 3 ," *Monthly Notices of*

the Royal Astronomical Society, vol. 421, no. 4, pp. 2809–2819, apr 2012. [Online]. Available: https://doi.org/10.1111/j.1365-2966.2012.20487.x [Cited on page 6.]

[40] A. Mesinger and Z. Haiman, "Evidence of a Cosmological Strömgren Surface and of Significant Neutral Hydrogen Surrounding the Quasar SDSS J1030+0524 ," *The Astrophysical Journal*, vol. 611, no. 2, p. L69–L72, jul 2004. [Online]. Available: http://dx.doi.org/10.1086/423935 [Cited on page 6.]

[41] D. A. Neufeld, "The Transfer of Resonance-Line Radiation in Static Astrophysical Media ," , vol. 350, p. 216, feb 1990. [Cited on page 6.]

[42] Verhamme, A., Schaerer, D., and Maselli, A., "3D Lyα radiation transfer - I. Understanding Lyα line profile morphologies ," *AA*, vol. 460, no. 2, pp. 397–413, 2006. [Online]. Available: https://doi.org/10.1051/0004-6361:20065554 [Cited on pages 6 and 7.]

[43] M. Hansen and S. P. Oh, "Lyman α radiative transfer in a multiphase medium ," *Monthly Notices of the Royal Astronomical Society*, vol. 367, no. 3, pp. 979–1002, apr 2006. [Online]. Available: https://doi.org/10.1111/j.1365-2966.2005.09870.x [Cited on page 6.]

[44] Atek, H., Kunth, D., Hayes, M., Östlin, G., and Mas-Hesse, J. M., "On the detectability of Lyα emission in star forming galaxies: The role of dust ," *AA*, vol. 488, no. 2, pp. 491–509, 2008. [Online]. Available: https://doi.org/10.1051/0004-6361:200809527 [Cited on page 7.]

[45] M. Dijkstra, A. Mesinger, and J. S. B. Wyithe, "The detectability of Lyα emission from galaxies during the epoch of reionization: The detectability of LAEs during the EoR ," *Monthly Notices of the Royal Astronomical Society*, vol. 414, no. 3, p. 2139–2147, apr 2011. [Online]. Available: http://dx.doi.org/10.1111/j.1365-2966.2011.18530.x [Cited on page 7.]

[46] M. R. Santos, "Probing reionization with Lyman α emission lines ," *Monthly Notices of the Royal Astronomical Society*, vol. 349, no. 3, pp. 1137–1152, apr 2004. [Online]. Available: https://doi.org/10.1111/j.1365-2966.2004.07594.x [Cited on page 7.]

[47] A. E. Shapley, C. C. Steidel, M. Pettini, and K. L. Adelberger, "Rest-Frame Ultraviolet Spectra of Lyman Break Galaxies at $z \sim 3$," *The Astrophysical*

Journal, vol. 588, no. 1, pp. 65–89, may 2003. [Online]. Available: http://dx.doi.org/10.1086/373922 [Cited on pages 7 and 8.]

[48] Atek, H., Schaerer, D., and Kunth, D., "Origin of Lyα absorption in nearby starbursts and implications for other galaxies ," *AA*, vol. 502, no. 3, pp. 791–801, 2009. [Online]. Available: https://doi.org/10.1051/0004-6361/200911856 [Cited on page 7.]

[49] D. A. Neufeld, "The Escape of Lyman-Alpha Radiation from a Multiphase Interstellar Medium ," , vol. 370, p. L85, apr 1991. [Cited on page 7.]

[50] Gronke, Max, Dijkstra, Mark, McCourt, Michael, and Peng Oh, S., "Resonant line transfer in a fog: using Lyman-alpha to probe tiny structures in atomic gas ," *AA*, vol. 607, p. A71, 2017. [Online]. Available: https://doi.org/10.1051/0004-6361/201731013 [Cited on page 7.]

[51] A. Mesinger and S. R. Furlanetto, "Lyα emitters during the early stages of reionization ," *Monthly Notices of the Royal Astronomical Society*, vol. 386, no. 4, p. 1990–2002, jun 2008. [Online]. Available: http://dx.doi.org/10.1111/j.1365-2966.2008.13039.x [Cited on page 7.]

[52] N. Laporte, K. Nakajima, R. S. Ellis, A. Zitrin, D. P. Stark, R. Mainali, and G. W. Roberts-Borsani, "A Spectroscopic Search for AGN Activity in the Reionization Era ," *The Astrophysical Journal*, vol. 851, no. 1, p. 40, dec 2017. [Online]. Available: https://dx.doi.org/10.3847/1538-4357/aa96a8

[53] M. McQuinn, A. Lidz, O. Zahn, S. Dutta, L. Hernquist, and M. Zaldarriaga, "The morphology of HII regions during reionization ," , vol. 377, no. 3, pp. 1043–1063, may 2007. [Cited on page 7.]

[54] D. Sobral, J. Matthee, B. Darvish, D. Schaerer, B. Mobasher, H. J. A. Röttgering, S. Santos, and S. Hemmati, "EVIDENCE FOR PopIII-LIKE STELLAR POPULATIONS IN THE MOST LUMINOUS LyαEMITTERS AT THE EPOCH OF REIONIZATION: SPECTROSCOPIC CONFIRMATION ," *The Astrophysical Journal*, vol. 808, no. 2, p. 139, jul 2015. [Online]. Available: http://dx.doi.org/10.1088/0004-637X/808/2/139 [Cited on page 7.]

[55] A. A. Khostovan, D. Sobral, B. Mobasher, J. Matthee, R. K. Cochrane, N. Chartab, M. Jafariyazani, A. Paulino-Afonso, S. Santos, and J. Calhau, "The clustering of typical Lyα emitters from $z \sim 2.5 - 6$: host halo masses depend on Lyα and UV luminosities ," , vol. 489, no. 1, pp. 555–573, oct 2019. [Cited on page 7.]

[56] Y. Harikane, M. Ouchi, T. Shibuya, T. Kojima, H. Zhang, R. Itoh, Y. Ono, R. Higuchi, A. K. Inoue, J. Chevallard, P. L. Capak, T. Nagao, M. Onodera, A. L. Faisst, C. L. Martin, M. Rauch, G. A. Bruzual, S. Charlot, I. Davidzon, S. Fujimoto, M. Hilmi, O. Ilbert, C.-H. Lee, Y. Matsuoka, J. D. Silverman, and S. Toft, "SILVERRUSH. V. Census of Lyα, [O III] λ5007, Hα, and [C II] 158 μm Line Emission with $\sim$1000 LAEs at $z = 4.9$–7.0 Revealed with Subaru/HSC," *The Astrophysical Journal*, vol. 859, no. 2, p. 84, may 2018. [Online]. Available: https://dx.doi.org/10.3847/1538-4357/aabd80 [Cited on page 7.]

[57] K. K. Nilsson, G. Östlin, P. Møller, O. Möller-Nilsson, C. Tapken, W. Freudling, and J. P. U. Fynbo, "The nature of $z \sim 2.3$ Lyman-α emitters ," , vol. 529, p. A9, may 2011. [Cited on page 7.]

[58] B. Margony, "The Sloan Digital Sky Survey ," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 357, no. 1750, p. 93–103, jan 1999. [Online]. Available: http://dx.doi.org/10.1098/rsta.1999.0316 [Cited on page 8.]

[59] K. C. Chambers, E. A. Magnier, N. Metcalfe, H. A. Flewelling, M. E. Huber, C. Z. Waters, L. Denneau, P. W. Draper, D. Farrow, D. P. Finkbeiner, C. Holmberg, J. Koppenhoefer, P. A. Price, A. Rest, R. P. Saglia, E. F. Schlafly, S. J. Smartt, W. Sweeney, R. J. Wainscoat, W. S. Burgett, S. Chastel, T. Grav, J. N. Heasley, K. W. Hodapp, R. Jedicke, N. Kaiser, R. P. Kudritzki, G. A. Luppino, R. H. Lupton, D. G. Monet, J. S. Morgan, P. M. Onaka, B. Shiao, C. W. Stubbs, J. L. Tonry, R. White, E. Bañados, E. F. Bell, R. Bender, E. J. Bernard, M. Boegner, F. Boffi, M. T. Botticella, A. Calamida, S. Casertano, W. P. Chen, X. Chen, S. Cole, N. Deacon, C. Frenk, A. Fitzsimmons, S. Gezari, V. Gibbs, C. Goessl, T. Goggia, R. Gourgue, B. Goldman, P. Grant, E. K. Grebel, N. C. Hambly, G. Hasinger, A. F. Heavens, T. M. Heckman, R. Henderson, T. Henning, M. Holman, U. Hopp, W. H. Ip, S. Isani, M. Jackson, C. D. Keyes, A. M. Koekemoer, R. Kotak, D. Le, D. Liska, K. S. Long, J. R. Lucey, M. Liu, N. F. Martin, G. Masci, B. McLean, E. Mindel, P. Misra, E. Morganson,

D. N. A. Murphy, A. Obaika, G. Narayan, M. A. Nieto-Santisteban, P. Norberg, J. A. Peacock, E. A. Pier, M. Postman, N. Primak, C. Rae, A. Rai, A. Riess, A. Riffeser, H. W. Rix, S. Röser, R. Russel, L. Rutz, E. Schilbach, A. S. B. Schultz, D. Scolnic, L. Strolger, A. Szalay, S. Seitz, E. Small, K. W. Smith, D. R. Soderblom, P. Taylor, R. Thomson, A. N. Taylor, A. R. Thakar, J. Thiel, D. Thilker, D. Unger, Y. Urata, J. Valenti, J. Wagner, T. Walder, F. Walter, S. P. Watters, S. Werner, W. M. Wood-Vasey, and R. Wyse, "The Pan-STARRS1 Surveys ," 2019. [Online]. Available: https://arxiv.org/abs/1612.05560 [Cited on page 8.]

[60] M. E. Levi, L. E. Allen, A. Raichoor, C. Baltay, S. BenZvi, F. Beutler, A. Bolton, F. J. Castander, C.-H. Chuang, A. Cooper, J.-G. Cuby, A. Dey, D. Eisenstein, X. Fan, B. Flaugher, C. Frenk, A. X. Gonzalez-Morales, O. Graur, J. Guy, S. Habib, K. Honscheid, S. Juneau, J.-P. Kneib, O. Lahav, D. Lang, A. Leauthaud, B. Lusso, A. de la Macorra, M. Manera, P. Martini, S. Mao, J. A. Newman, N. Palanque-Delabrouille, W. J. Percival, C. A. Prieto, C. M. Rockosi, V. Ruhlmann-Kleider, D. Schlegel, H.-J. Seo, Y.-S. Song, G. Tarle, R. Wechsler, D. Weinberg, C. Yeche, and Y. Zu, "The Dark Energy Spectroscopic Instrument (DESI) ," 2019. [Online]. Available: https://arxiv.org/abs/1907.10688 [Cited on page 8.]

[61] K. Breivik, A. J. Connolly, K. E. S. Ford, M. Jurić, R. Mandelbaum, A. A. Miller, D. Norman, K. Olsen, W. O'Mullane, A. Price-Whelan, T. Sacco, J. L. Sokoloski, A. Villar, V. Acquaviva, T. Ahumada, Y. AlSayyad, C. S. Alves, I. Andreoni, T. Anguita, H. J. Best, F. B. Bianco, R. Bonito, A. Bradshaw, C. J. Burke, A. R. de Campos, M. Cantiello, N. Caplar, C. O. Chandler, J. Chan, L. N. da Costa, S. Danieli, J. R. A. Davenport, G. Fabbian, J. Fagin, A. Gagliano, C. Gall, N. G. Camargo, E. Gawiser, S. Gezari, A. Gomboc, A. X. Gonzalez-Morales, M. J. Graham, J. Gschwend, L. P. Guy, M. J. Holman, H. H. Hsieh, M. Hundertmark, D. Ilić, E. E. O. Ishida, T. Jurkić, A. Kannawadi, A. Kosakowski, A. B. Kovačević, J. Kubica, F. Lanusse, I. Lazar, W. G. Levine, X. Li, J. Lu, G. J. M. Luna, A. A. Mahabal, A. I. Malz, Y.-Y. Mao, I. Medan, J. Moeyens, M. Nikolić, R. Nikutta, M. O'Dowd, C. Olsen, S. Pearson, I. V. Pedraza, M. Popinchalk, L. C. Popović, T. A. Pritchard, B. C. Quint, V. Radović, F. Ragosta, G. Riccio, A. H. Riley, A. Rożek, P. Sánchez-Sáez, L. M. Sarro, C. Saunders, Đorđe V. Savić, S. Schmidt, A. Scott, R. Shirley, H. R. Smotherman, S. Stetzler, K. Storey-Fisher, R. A. Street, D. E. Trilling, Y. Tsapras, S. Ustamujic, S. van Velzen, J. A. Vázquez-Mata, L. Venuti, S. Wyatt, W. Yu, and

A. Zabludoff, "From Data to Software to Science with the Rubin Observatory LSST ," 2022. [Online]. Available: https://arxiv.org/abs/2208.02781 [Cited on page 8.]

[62] J. Lazio, "The Square Kilometre Array ," 2009. [Online]. Available: https://arxiv.org/abs/0910.0632 [Cited on page 8.]

[63] A. M. Koekemoer, H. Aussel, D. Calzetti, P. Capak, M. Giavalisco, J. Kneib, A. Leauthaud, O. Le Fevre, H. J. McCracken, R. Massey, B. Mobasher, J. Rhodes, N. Scoville, and P. L. Shopbell, "The COSMOS Survey: Hubble Space Telescope Advanced Camera for Surveys Observations and Data Processing," *The Astrophysical Journal Supplement Series*, vol. 172, no. 1, p. 196–202, sep 2007. [Online]. Available: http://dx.doi.org/10.1086/520086 [Cited on page 8.]

[64] J. R. Weaver, O. B. Kauffmann, O. Ilbert, H. J. McCracken, A. Moneti, S. Toft, G. Brammer, M. Shuntov, I. Davidzon, B. C. Hsieh, C. Laigle, A. Anastasiou, C. K. Jespersen, J. Vinther, P. Capak, C. M. Casey, C. J. R. McPartland, B. Milvang-Jensen, B. Mobasher, D. B. Sanders, L. Zalesky, S. Arnouts, H. Aussel, J. S. Dunlop, A. Faisst, M. Franx, L. J. Furtak, J. P. U. Fynbo, K. M. L. Gould, T. R. Greve, S. Gwyn, J. S. Kartaltepe, D. Kashino, A. M. Koekemoer, V. Kokorev, O. Le Fèvre, S. Lilly, D. Masters, G. Magdis, V. Mehta, Y. Peng, D. A. Riechers, M. Salvato, M. Sawicki, C. Scarlata, N. Scoville, R. Shirley, J. D. Silverman, A. Sneppen, V. Smolčić, C. Steinhardt, D. Stern, M. Tanaka, Y. Taniguchi, H. I. Teplitz, M. Vaccari, W.-H. Wang, and G. Zamorani, "COSMOS2020: A Panchromatic View of the Universe to $z \sim 10$ from Two Complementary Catalogs ," *The Astrophysical Journal Supplement Series*, vol. 258, no. 1, p. 11, jan 2022. [Online]. Available: http://dx.doi.org/10.3847/1538-4365/ac3078 [Cited on pages 8, 13, 14, and 17.]

[65] C. G. Díaz, E. V. Ryan-Weber, W. Karman, K. I. Caputi, S. Salvadori, N. H. Crighton, M. Ouchi, and E. Vanzella, "Faint LAEs near z4.7 CIV absorbers revealed by MUSE ," *Monthly Notices of the Royal Astronomical Society*, vol. 502, no. 2, p. 2645–2663, oct 2020. [Online]. Available: http://dx.doi.org/10.1093/mnras/staa3129 [Cited on page 8.]

[66] A. Saxena, B. E. Robertson, A. J. Bunker, R. Endsley, A. J. Cameron, S. Charlot, C. Simmonds, S. Tacchella, J. Witstok, C. Willott, S. Carniani, E. Curtis-Lake, P. Ferruit, P. Jakobsen, S. Arribas, J. Chevallard, M. Curti, F. D'Eugenio,

A. De Graaff, G. C. Jones, T. J. Looser, M. V. Maseda, T. Rawle, H.-W. Rix, B. R. Del Pino, R. Smit, H. Übler, D. J. Eisenstein, K. Hainline, R. Hausen, B. D. Johnson, M. Rieke, C. C. Williams, C. N. A. Willmer, W. M. Baker, R. Bhatawdekar, R. Bowler, K. Boyett, Z. Chen, E. Egami, Z. Ji, N. Kumari, E. Nelson, M. Perna, L. Sandles, J. Scholtz, and I. Shivaei, "JADES: Discovery of extremely high equivalent width Lyman-α emission from a faint galaxy within an ionized bubble at $z = 7.3$," *Astronomy amp; Astrophysics*, vol. 678, p. A68, oct 2023. [Online]. Available: http://dx.doi.org/10.1051/0004-6361/202346245

[67] A. J. Bunker, A. Saxena, A. J. Cameron, C. J. Willott, E. Curtis-Lake, P. Jakobsen, S. Carniani, R. Smit, R. Maiolino, J. Witstok, M. Curti, F. D'Eugenio, G. C. Jones, P. Ferruit, S. Arribas, S. Charlot, J. Chevallard, G. Giardino, A. de Graaff, T. J. Looser, N. Lützgendorf, M. V. Maseda, T. Rawle, H.-W. Rix, B. R. Del Pino, S. Alberts, E. Egami, D. J. Eisenstein, R. Endsley, K. Hainline, R. Hausen, B. D. Johnson, G. Rieke, M. Rieke, B. E. Robertson, I. Shivaei, D. P. Stark, F. Sun, S. Tacchella, M. Tang, C. C. Williams, C. N. A. Willmer, W. M. Baker, S. Baum, R. Bhatawdekar, R. Bowler, K. Boyett, Z. Chen, C. Circosta, J. M. Helton, Z. Ji, N. Kumari, J. Lyu, E. Nelson, E. Parlanti, M. Perna, L. Sandles, J. Scholtz, K. A. Suess, M. W. Topping, H. Übler, I. E. B. Wallace, and L. Whitler, "JADES NIRSpec Spectroscopy of GN-z11: Lyman-α emission and possible enhanced nitrogen abundance in a $z = 10.60$ luminous galaxy ," *Astronomy amp; Astrophysics*, vol. 677, p. A88, sep 2023. [Online]. Available: http://dx.doi.org/10.1051/0004-6361/202346159 [Cited on page 8.]

[68] N. M. Firestone, E. Gawiser, V. Ramakrishnan, K.-S. Lee, F. Valdes, C. Park, Y. Yang, R. Ciardullo, M. C. Artale, B. Benda, A. Broussard, L. Eid, R. Farooq, C. Gronwall, L. Guaita, S. Gwyn, H. S. Hwang, S. H. Im, W.-S. Jeong, S. Karthikeyan, D. Lang, B. Moon, N. Padilla, M. Sawicki, E. Seo, A. Singh, H. Song, and P. Troncoso Iribarren, "ODIN: Improved Narrowband Lyα Emitter Selection Techniques for $z = 2.4, 3.1, and 4.5$," *The Astrophysical Journal*, vol. 974, no. 2, p. 217, oct 2024. [Online]. Available: http://dx.doi.org/10.3847/1538-4357/ad71c9 [Cited on page 8.]

[69] L. Pozzetti, M. Bolzonella, E. Zucca, G. Zamorani, S. Lilly, A. Renzini, M. Moresco, M. Mignoli, P. Cassata, L. Tasca, F. Lamareille, C. Maier, B. Meneux, C. Halliday,

P. Oesch, D. Vergani, K. Caputi, K. Kovač, A. Cimatti, O. Cucciati, A. Iovino, Y. Peng, M. Carollo, T. Contini, J.-P. Kneib, O. Le Févre, V. Mainieri, M. Scodeggio, S. Bardelli, A. Bongiorno, G. Coppa, S. de la Torre, L. de Ravel, P. Franzetti, B. Garilli, P. Kampczyk, C. Knobel, J.-F. Le Borgne, V. Le Brun, R. Pellò, E. Perez Montero, E. Ricciardelli, J. D. Silverman, M. Tanaka, L. Tresse, U. Abbas, D. Bottini, A. Cappi, L. Guzzo, A. M. Koekemoer, A. Leauthaud, D. Maccagni, C. Marinoni, H. J. McCracken, P. Memeo, C. Porciani, R. Scaramella, C. Scarlata, and N. Scoville, "zCOSMOS – 10k-bright spectroscopic sample: The bimodality in the galaxy stellar mass function: exploring its evolution with redshift ," *Astronomy amp; Astrophysics*, vol. 523, p. A13, nov 2010. [Online]. Available: http://dx.doi.org/10.1051/0004-6361/200913020 [Cited on page 8.]

[70] G. Hasinger, P. Capak, M. Salvato, A. J. Barger, L. L. Cowie, A. Faisst, S. Hemmati, Y. Kakazu, J. Kartaltepe, D. Masters, B. Mobasher, H. Nayyeri, D. Sanders, N. Z. Scoville, H. Suh, C. Steinhardt, and F. Yang, "The DEIMOS 10K Spectroscopic Survey Catalog of the COSMOS Field ," *The Astrophysical Journal*, vol. 858, no. 2, p. 77, may 2018. [Online]. Available: http://dx.doi.org/10.3847/1538-4357/aabacf [Cited on page 8.]

[71] L. A. M. Tasca, O. Le Fèvre, B. Ribeiro, R. Thomas, C. Moreau, P. Cassata, B. Garilli, V. Le Brun, B. C. Lemaux, D. Maccagni, L. Pentericci, D. Schaerer, E. Vanzella, G. Zamorani, E. Zucca, R. Amorin, S. Bardelli, L. P. Cassarà, M. Castellano, A. Cimatti, O. Cucciati, A. Durkalec, A. Fontana, M. Giavalisco, A. Grazian, N. P. Hathi, O. Ilbert, S. Paltani, J. Pforr, M. Scodeggio, V. Sommariva, M. Talia, L. Tresse, D. Vergani, P. Capak, S. Charlot, T. Contini, S. de la Torre, J. Dunlop, S. Fotopoulou, L. Guaita, A. Koekemoer, C. López-Sanjuan, Y. Mellier, M. Salvato, N. Scoville, Y. Taniguchi, and P. W. Wang, "The VIMOS Ultra Deep Survey first data release: Spectra and spectroscopic redshifts of 698 objects up to $zspec \approx 6$ in CANDELS ," *Astronomy amp; Astrophysics*, vol. 600, p. A110, apr 2017. [Online]. Available: http://dx.doi.org/10.1051/0004-6361/201527963 [Cited on page 8.]

[72] R. Bacon, S. Conseil, D. Mary, J. Brinchmann, M. Shepherd, M. Akhlaghi, P. M. Weilbacher, L. Piqueras, L. Wisotzki, D. Lagattuta, B. Epinat, A. Guerou, H. Inami, S. Cantalupo, J. B. Courbot, T. Contini, J. Richard, M. Maseda, R. Bouwens, N. Bouché, W. Kollatschny, J. Schaye, R. A. Marino, R. Pello,

C. Herenz, B. Guiderdoni, and M. Carollo, "The MUSE Hubble Ultra Deep Field Survey: I. Survey description, data reduction, and source detection ," *Astronomy amp; Astrophysics*, vol. 608, p. A1, nov 2017. [Online]. Available: http://dx.doi.org/10.1051/0004-6361/201730833 [Cited on page 8.]

[73] X. Fan, M. A. Strauss, D. P. Schneider, J. E. Gunn, R. H. Lupton, B. Yanny, S. F. Anderson, J. E. Anderson, Jr., J. Annis, N. A. Bahcall, J. A. Bakken, S. Bastian, E. Berman, W. N. Boroski, C. Briegel, J. W. Briggs, J. Brinkmann, M. A. Carr, P. L. Colestock, A. J. Connolly, J. H. Crocker, I. Csabai, P. C. Czarapata, J. E. Davis, M. Doi, B. R. Elms, M. L. Evans, G. R. Federwitz, J. A. Frieman, M. Fukugita, V. K. Gurbani, F. H. Harris, T. M. Heckman, G. S. Hennessy, R. B. Hindsley, D. J. Holmgren, C. Hull, S.-I. Ichikawa, T. Ichikawa, Ivezić, S. Kent, G. R. Knapp, R. G. Kron, D. Q. Lamb, R. F. Leger, S. Limmongkol, C. Lindenmeyer, D. C. Long, J. Loveday, B. MacKinnon, E. J. Mannery, P. M. Mantsch, B. Margon, T. A. McKay, J. A. Munn, T. Nash, H. J. Newberg, R. C. Nichol, T. Nicinski, S. Okamura, J. P. Ostriker, R. Owen, A. G. Pauls, J. Peoples, D. Petravick, J. R. Pier, R. Pordes, A. Prosapio, R. Rechenmacher, G. T. Richards, M. W. Richmond, C. H. Rivetta, C. M. Rockosi, D. Sandford, G. Sergey, M. Sekiguchi, K. Shimasaku, W. A. Siegmund, J. A. Smith, C. Stoughton, A. S. Szalay, G. P. Szokoly, D. L. Tucker, M. S. Vogeley, P. Waddell, S.-i. Wang, D. H. Weinberg, N. Yasuda, and D. G. York, "High-Redshift Quasars Found in Sloan Digital Sky Survey Commissioning Data ," *The Astronomical Journal*, vol. 118, no. 1, p. 1–13, jul 1999. [Online]. Available: http://dx.doi.org/10.1086/300944 [Cited on page 8.]

[74] M. Yoshida, K. Shimasaku, M. Ouchi, K. Sekiguchi, H. Furusawa, and S. Okamura, "The Subaru/XMM-NewtonDeep Survey (SXDS). VII. Clustering Segregation with Ultraviolet and Optical Luminosities of Lyman Break Galaxies at $z \approx 3$," *The Astrophysical Journal*, vol. 679, no. 1, p. 269–278, may 2008. [Online]. Available: http://dx.doi.org/10.1086/586726 [Cited on page 8.]

[75] C. Gronwall, R. Ciardullo, T. Hickey, E. Gawiser, J. J. Feldmeier, P. G. van Dokkum, C. M. Urry, D. Herrera, B. D. Lehmer, L. Infante, A. Orsi, D. Marchesini, G. A. Blanc, H. Francke, P. Lira, and E. Treister, "Lyα Emission-Line Galaxies at $z = 3.1$ in the Extended Chandra Deep Field–South ," *The Astrophysical Journal*, vol. 667, no. 1,

p. 79–91, sep 2007. [Online]. Available: http://dx.doi.org/10.1086/520324 [Cited on page 8.]

[76] T. Yoshioka, N. Kashikawa, Y. Takeda, K. Ito, Y. Liang, R. Ishimoto, J. Arita, Y. Nishimura, H. Hoshi, and S. Shimizu, "Predicting Lyα Emission from Distant Galaxies with Neural Network Architecture ," 2024. [Online]. Available: https://arxiv.org/abs/2412.14676 [Cited on page 9.]

[77] J. B. Oke and J. E. Gunn, "Secondary standard stars for absolute spectrophotometry ," , vol. 266, pp. 713–717, mar 1983. [Cited on page 11.]

[78] H. Aihara, N. Arimoto, R. Armstrong, S. Arnouts, N. A. Bahcall, S. Bickerton, J. Bosch, K. Bundy, P. L. Capak, J. H. H. Chan, M. Chiba, J. Coupon, E. Egami, M. Enoki, F. Finet, H. Fujimori, S. Fujimoto, H. Furusawa, J. Furusawa, T. Goto, A. Goulding, J. P. Greco, J. E. Greene, J. E. Gunn, T. Hamana, Y. Harikane, Y. Hashimoto, T. Hattori, M. Hayashi, Y. Hayashi, K. G. Hełminiak, R. Higuchi, C. Hikage, P. T. P. Ho, B.-C. Hsieh, K. Huang, S. Huang, H. Ikeda, M. Imanishi, A. K. Inoue, K. Iwasawa, I. Iwata, A. T. Jaelani, H.-Y. Jian, Y. Kamata, H. Karoji, N. Kashikawa, N. Katayama, S. Kawanomoto, I. Kayo, J. Koda, M. Koike, T. Kojima, Y. Komiyama, A. Konno, S. Koshida, Y. Koyama, H. Kusakabe, A. Leauthaud, C.-H. Lee, L. Lin, Y.-T. Lin, R. H. Lupton, R. Mandelbaum, Y. Matsuoka, E. Medezinski, S. Mineo, S. Miyama, H. Miyatake, S. Miyazaki, R. Momose, A. More, S. More, Y. Moritani, T. J. Moriya, T. Morokuma, S. Mukae, R. Murata, H. Murayama, T. Nagao, F. Nakata, M. Niida, H. Niikura, A. J. Nishizawa, Y. Obuchi, M. Oguri, Y. Oishi, N. Okabe, S. Okamoto, Y. Okura, Y. Ono, M. Onodera, M. Onoue, K. Osato, M. Ouchi, P. A. Price, T.-S. Pyo, M. Sako, M. Sawicki, T. Shibuya, K. Shimasaku, A. Shimono, M. Shirasaki, J. D. Silverman, M. Simet, J. Speagle, D. N. Spergel, M. A. Strauss, Y. Sugahara, N. Sugiyama, Y. Suto, S. H. Suyu, N. Suzuki, P. J. Tait, M. Takada, T. Takata, N. Tamura, M. M. Tanaka, M. Tanaka, M. Tanaka, Y. Tanaka, T. Terai, Y. Terashima, Y. Toba, N. Tominaga, J. Toshikawa, E. L. Turner, T. Uchida, H. Uchiyama, K. Umetsu, F. Uraguchi, Y. Urata, T. Usuda, Y. Utsumi, S.-Y. Wang, W.-H. Wang, K. C. Wong, K. Yabe, Y. Yamada, H. Yamanoi, N. Yasuda, S. Yeh, A. Yonehara, and S. Yuma, "The Hyper Suprime-Cam SSP Survey: Overview and survey design ," *Publications of the Astronomical Society of Japan*, vol. 70, no. SP1,

sep 2017. [Online]. Available: http://dx.doi.org/10.1093/pasj/psx066 [Cited on pages 13 and 22.]

[79] J. J. A. Matthee, D. Sobral, A. M. Swinbank, I. Smail, P. N. Best, J.-W. Kim, M. Franx, B. Milvang-Jensen, and J. Fynbo, "A 10 deg^2 Lyman α survey at $z = 8.8$ with spectroscopic follow-up: strong constraints on the luminosity function and implications for other surveys ," *Monthly Notices of the Royal Astronomical Society*, vol. 440, no. 3, pp. 2375–2387, mar 2014. [Online]. Available: http://dx.doi.org/10.1093/mnras/stu392 [Cited on page 15.]

[80] M. Taylor, "TOPCAT: Working with Data and Working with Users ," 2017. [Cited on page 17.]

[81] J. Calhau, D. Sobral, S. Santos, J. Matthee, A. Paulino-Afonso, A. Stroe, B. Simmons, and C. Barlow-Hall, "The X-ray and radio activity of typical and luminous Lya emitters from $z \sim 2$ to $z \sim 6$: Evidence for a diverse, evolving population ," *Monthly Notices of the Royal Astronomical Society*, vol. 493, no. 3, pp. 3341–3362, apr 2020. [Cited on page 18.]

[82] P. Capak, H. Aussel, M. Ajiki, H. J. McCracken, B. Mobasher, N. Scoville, P. Shopbell, Y. Taniguchi, D. Thompson, S. Tribiano, S. Sasaki, A. W. Blain, M. Brusa, C. Carilli, A. Comastri, C. M. Carollo, P. Cassata, J. Colbert, R. S. Ellis, M. Elvis, M. Giavalisco, W. Green, L. Guzzo, G. Hasinger, O. Ilbert, C. Impey, K. Jahnke, J. Kartaltepe, J. P. Kneib, J. Koda, A. Koekemoer, Y. Komiyama, A. Leauthaud, O. Le Fevre, S. Lilly, C. Liu, R. Massey, S. Miyazaki, T. Murayama, T. Nagao, J. A. Peacock, A. Pickles, C. Porciani, A. Renzini, J. Rhodes, M. Rich, M. Salvato, D. B. Sanders, C. Scarlata, D. Schiminovich, E. Schinnerer, M. Scodeggio, K. Sheth, Y. Shioya, L. A. M. Tasca, J. E. Taylor, L. Yan, and G. Zamorani, " The First Release COSMOS Optical and Near-IR Data and Catalog ," *The Astrophysical Journal Supplement Series*, vol. 172, no. 1, pp. 99–116, sep 2007. [Cited on page 18.]

[83] C. Laigle, H. J. McCracken, O. Ilbert, B. C. Hsieh, I. Davidzon, P. Capak, G. Hasinger, J. D. Silverman, C. Pichon, J. Coupon, H. Aussel, D. Le Borgne, K. Caputi, P. Cassata, Y. Y. Chang, F. Civano, J. Dunlop, J. Fynbo, J. S. Kartaltepe, A. Koekemoer, O. Le Fèvre, E. Le Floc'h, A. Leauthaud, S. Lilly, L. Lin, S. Marchesi, B. Milvang-Jensen, M. Salvato, D. B. Sanders, N. Scoville, V. Smolcic, M. Stockmann,

Y. Taniguchi, L. Tasca, S. Toft, M. Vaccari, and J. Zabl, "The COSMOS2015 Catalog: Exploring the $1 < z < 6$ Universe with Half a Million Galaxies ," *The Astrophysical Journal Supplement Series*, vol. 224, no. 2, p. 24, jun 2016. [Cited on page 18.]

[84] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python ," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Cited on page 18.]

[85] A. Collaboration, T. P. Robitaille, E. J. Tollerud, P. Greenfield, M. Droettboom, E. Bray, T. Aldcroft, M. Davis, A. Ginsburg, A. M. Price-Whelan, W. E. Kerzendorf, A. Conley, N. Crighton, K. Barbary, D. Muna, H. Ferguson, F. Grollier, M. M. Parikh, P. H. Nair, H. M. Unther, C. Deil, J. Woillez, S. Conseil, R. Kramer, J. E. H. Turner, L. Singer, R. Fox, B. A. Weaver, V. Zabalza, Z. I. Edwards, K. Azalee Bostroem, D. J. Burke, A. R. Casey, S. M. Crawford, N. Dencheva, J. Ely, T. Jenness, K. Labrie, P. L. Lim, F. Pierfederici, A. Pontzen, A. Ptak, B. Refsdal, M. Servillat, and O. Streicher, "Astropy: A community Python package for astronomy ," *A&A*, vol. 558, p. A33, jan 2013. [Cited on page 22.]

[86] R. Lupton, M. R. Blanton, G. Fekete, D. W. Hogg, W. O'Mullane, A. Szalay, and N. Wherry, "Preparing Red-Green-Blue Images from CCD Data ," , vol. 116, no. 816, pp. 133–137, feb 2004. [Cited on page 22.]

[87] M. Awad and R. Khanna, *Machine Learning* . Berkeley, CA: Apress, 2015, pp. 1–18. [Online]. Available: https://doi.org/10.1007/978-1-4302-5990-9_1 [Cited on page 24.]

[88] H. Sheikh, C. Prins, and E. Schrijvers, *Artificial Intelligence: Definition and Background* . Cham: Springer International Publishing, 2023, pp. 15–41. [Online]. Available: https://doi.org/10.1007/978-3-031-21448-6_2 [Cited on page 24.]

[89] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks ," *Science*, vol. 313, no. 5786, pp. 504–507, 2006. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1127647 [Cited on page 24.]

[90] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition ," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. [Cited on page 24.]

[91] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-dujaili, Y. Duan, O. Al-Shamma, J. I. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232434552 [Cited on page 25.]

[92] H. J. Kelley, "Gradient Theory of Optimal Flight Paths ," *ARS Journal*, vol. 30, pp. 947–954, 1960. [Online]. Available: https://api.semanticscholar.org/CorpusID:121072881 [Cited on page 26.]

[93] P. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral science. thesis (ph. d.). appl. math. harvard university," Ph.D. dissertation, Harvard University, jan 1974. [Cited on page 26.]

[94] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines ," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. Madison, WI, USA: Omnipress, 2010, p. 807–814. [Cited on page 27.]

[95] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors ," *Nature*, vol. 323, pp. 533–536, 1986. [Online]. Available: https://api.semanticscholar.org/CorpusID:205001834 [Cited on page 27.]

[96] J. S. Bridle, "Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition ," in *Neurocomputing*, F. F. Soulié and J. Hérault, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1990, pp. 227–236. [Cited on page 28.]

[97] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005. [Online]. Available: http://www.jstor.org/stable/24869236 [Cited on page 30.]

[98] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014. [Online]. Available: https://gmd.copernicus.org/articles/7/1247/2014/ [Cited on page 31.]

[99] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy ," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207006000239 [Cited on page 31.]

[100] C. Cortes, M. Mohri, and A. Rostamizadeh, "L2 regularization for learning kernels," 2012. [Online]. Available: https://arxiv.org/abs/1205.2653 [Cited on page 33.]

[101] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi *et al.*, "KerasTuner," https://github.com/keras-team/keras-tuner, 2019. [Cited on page 35.]

[102] F. Chollet *et al.*, "Keras ," https://keras.io, 2015, accessed: 2025-05-02. [Cited on page 36.]

[103] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1409.1556 [Cited on page 37.]

[104] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807. [Cited on page 37.]

[105] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385 [Cited on page 37.]

[106] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018. [Online]. Available: https://arxiv.org/abs/1608.06993 [Cited on page 37.]

[107] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929 [Cited on page 37.]

[108] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: removing noise by adding noise ," 2017. [Online]. Available: https://arxiv.org/abs/1706.03825 [Cited on page 39.]

[109] E. Mentuch Cooper, K. Gebhardt, D. Davis, D. J. Farrow, C. Liu, G. Zeimann, R. Ciardullo, J. J. Feldmeier, N. Drory, D. Jeong, B. Benda, W. P. Bowman, M. Boylan-Kolchin, Ó. A. Chávez Ortiz, M. H. Debski, M. Dentler, M. Fabricius, R. Farooq, S. L. Finkelstein, E. Gawiser, C. Gronwall, G. J. Hill, U. Hopp, L. R. House, S. Janowiecki, H. Khoraminezhad, W. Kollatschny, E. Komatsu, M. Landriau, M. L. Niemeyer, H. Lee, P. MacQueen, K. Mawatari, B. McKay, M. Ouchi, J. Poppe, S. Saito, D. P. Schneider, J. Snigula, B. P. Thomas, S. Tuttle, T. Urrutia, L. Weiss, L. Wisotzki, Y. Zhang, and HETDEX Collaboration, "HETDEX Public Source Catalog 1: 220 K Sources Including Over 50 K Lyα Emitters from an Untargeted Wide-area Spectroscopic Survey ," , vol. 943, no. 2, p. 177, feb 2023. [Cited on pages ix, 51, 52, 73, and 77.]

[110] A. Paulino-Afonso, D. Sobral, B. Ribeiro, J. Matthee, S. Santos, J. Calhau, A. Forshaw, A. Johnson, J. Merrick, S. Pérez, and O. Sheldon, "On the UV compactness and morphologies of typical Lyman α emitters from $z \sim 2$ to ~ 6 ," *Monthly Notices of the Royal Astronomical Society*, vol. 476, no. 4, p. 5479–5501, feb 2018. [Online]. Available: http://dx.doi.org/10.1093/mnras/sty281 [Cited on page 84.]