Astronomy & Astrophysics manuscript no. aa55170-25corr
September 30, 2025

A gradient boosting and broadband approach to finding Lyman-α emitting galaxies beyond narrowband surveys

A. Vale[1,2], A. Paulino-Afonso[1], A. Humphrey[1,3], P.A.C. Cunha[1,2], B. Ribeiro[4], B. Cerqueira[1,2], R. Carvajal[5,6], and J. Fonseca[1,2]

[1] Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, PT4150-762 Porto, Portugal
[2] Departamento de Física e Astronomia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre 687, PT4169-007 Porto, Portugal
[3] DTx–Digital Transformation CoLab, Building 1, Azurém Campus, University of Minho, PT4800-058 Guimarães, Portugal
[4] Celfocus, Avenida Dom João II, 34, Parque das Nações, 1998-031, Lisbon, Portugal
[5] Instituto de Astrofísica e Ciências do Espaço, Universidade de Lisboa, OAL, Tapada da Ajuda, 1349-018 Lisbon, Portugal
[6] Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Edifício C8, Campo Grande, 1749-016 Lisbon, Portugal

September 30, 2025

ABSTRACT

Context. The identification of Lyman-α emitting galaxies (LAEs) has traditionally relied on dedicated surveys using custom narrowband filters, which constrain observations to specific narrow redshift intervals, or on blind spectroscopy, which although unbiased, typically requires extensive telescope time. This makes it challenging to assemble large statistically robust galaxy samples. With the advent of wide-area astronomical surveys producing datasets that are significantly larger than traditional surveys, the need for new techniques arises.

Aims. We test whether gradient-boosting algorithms, trained on broadband photometric data from traditional LAE surveys, can efficiently and accurately identify LAE candidates from typical star-forming galaxies at similar redshifts and brightness levels.

Methods. Using galaxy samples at $z \in [2, 6]$ derived from the COSMOS2020 and SC4K catalogs, we trained gradient-boosting machine-learning algorithms (LGBM, XGBoost, and CatBoost) using optical and near-infrared broadband photometry. To ensure balanced performance, the models were trained on carefully selected datasets with similar redshift and i-band magnitude distributions. Additionally, the models were tested for robustness by perturbing the photometric data using the associated observational uncertainties.

Results. Our classification models achieved F1-scores of $\sim 87\%$ and successfully identified about $7,000$ objects with an unanimous agreement across all models. This more than doubles the number of LAEs identified in the COSMOS field compared with the SC4K dataset. We managed to spectroscopically confirm 60 of these LAE candidates using the publicly available catalogs in the COSMOS field.

Conclusions. These results highlight the potential of machine learning in efficiently identifying LAEs candidates. This lays the foundations for applications to larger photometric surveys, such as Euclid and LSST. By complementing traditional approaches and providing robust preselection capabilities, our models facilitate the analysis of these objects. This is crucial to increase our knowledge of the overall LAE population.

Key words. galaxies: high redshift – galaxies: photometry – surveys – methods: data analysis – methods: statistical

. Introduction

The Lyman-α (hereafter, Lyα) emission line is one of the most prominent spectral features emitted by star-forming galaxies (e.g. Partridge & Peebles 1967; Ouchi et al. 2003; Dijkstra 014). With its short wavelength (1216Å) and intrinsic brightness, the Lyα emission proves highly effective for determining faint high-redshift objects through optical and near-infrared (NIR) observations (e.g. Ouchi et al. 2003; Sobral et al. 2018a; Harikane et al. 2018). Lyα emitters (hereafter referred to as LAEs) are galaxies with a detected Lyα emission line with a rest-frame equivalent width (EW_0) that exceeds approximately 20Å (e.g. Shibuya et al. 2019; Ouchi et al. 2020). LAEs are compact galaxies ($r_e \sim 1$ kpc, e.g. Paulino-Afonso et al. 2018), generally with a low stellar mass ($M_* \sim 10^{8-9} M_\odot$, e.g. Pentericci et al. 2007) and low metallicities ($Z < 2 \times 10^{-2} Z_\odot$, e.g. Ouchi et al. 2020). They exhibit an inverse correlation between galaxy size and Lyα emission, where smaller galaxies tend to have larger escape fractions (Law et al. 2012; Paulino-Afonso et al. 2018). Additionally, they present star formation rates of $\sim 1 - 10$ M$_\odot$ yr^{-1} (e.g. Gawiser et al. 2007) and young stellar ages (~ 10 Myr, e.g. Hagen et al. 2014; Nakajima et al. 2012). These properties suggest that LAEs trace the early galaxy evolution (Taniguchi et al. 2003).

The most common method for detecting LAEs is narrowband (NB) imaging (e.g. Hu et al. 2010; Matthee et al. 2015a), which involves comparing images taken using a NB filter (typically 100–200Å wide) (e.g. Ajiki et al. 2003; Grove et al. 2009) with a broadband image to identify a flux excess in the NB image. This technique is limited, however, because it only probes specific narrow redshift ranges and is biased toward higher equivalent widths (e.g. Matthee et al. 2015b). Furthermore, there is a non-negligible possibility that these NB-selected samples may include galaxies at lower redshifts with additional emission lines that may be confused with Lyα, such as CIV emission at 1549Å

(e.g. Fynbo et al. 2003), MgII at 2798Å (e.g. Dunlop 2013), [OII] at 3727Å (e.g. Fujita et al. 2003), or [OIII] at 5007Å (e.g. Ciardullo et al. 2002). Blind spectroscopy (e.g. Kurk et al. 2004) is another technique that is used to search for LAEs. Candidates are identified based on their emission lines without prior photometric selection. While it allows for the detection of LAEs without selection biases, its efficiency is limited by the fact that it often requires extensive telescope time. Integral field spectrographs (IFS) can also be used (e.g. Drake et al. 2017). They offer the advantage of capturing the spectra of unbiased galaxy samples. Their effectiveness as wide-area survey instruments is limited by their relatively small field of view, however.

Despite the success of these techniques, they each have different limitations, and as LAE searches expand with upcoming surveys, such as the Large Synoptic Survey Telescope (LSST; Ivezić et al. 2019) and Euclid (Euclid Collaboration et al. 2025), machine-learning methods will be essential to optimize the sample selection and mitigate biases (e.g. Huertas-Company & Lanusse 2023). Although the application of machine learning to LAE studies remains limited, some key works have been conducted. Runnholm et al. (2020) used a linear regression to predict the Lyα luminosity based upon observed properties such as broadband luminosities or other nebular lines. They also used derived physical properties, such as the stellar mass or star formation rate (SFR). Ono et al. (2021) trained convolutional neural network models using simulated images of LAEs to remove contaminants from photometrically selected LAE candidates by performing an image classification based on NB and broadband filter data. Finally, Napolitano et al. (2023) developed a machine-learning model based on random forest to select LAEs. They achieved an accuracy of (80 ± 2)% and a precision of (73 ± 4)% in the redshift range $z \in [2.5, 4.5]$. At higher redshifts ($z \in [4.5, 6]$), they obtained an accuracy of 73% and a precision of 80%. This model was built on the basis of the physical (stellar mass, SFR, reddening, metallicity, and age) and morphological properties (Sérsic index, half-light radius, and projected semimajor axis), and the training sample was based on several samples of spectroscopically confirmed LAEs from the Great Observatories Origins Deep Survey South field (GOODS-S; Giavalisco et al. 2004), Ultra Deep Survey (UDS; Lawrence et al. 2007), and Cosmic Evolution Survey Deep (COSMOS; Capak et al. 2007).

Instead of using derived physical properties, we aim to identify LAEs in a more efficient and cost-effective manner by using only broadband photometric data. We trained supervised machine-learning gradient-boosting algorithms on about 4,000 LAEs from the SC4K survey (Sobral et al. 2018b) and non-LAEs (nLAEs) from COSMOS2020 (Weaver et al. 2022) at $z \in [2, 6]$. We also built on correlations between LAEs and their fluxes, magnitudes, and colors in the optical and NIR broadband information to pave the way for a future application to larger surveys. This is valuable for surveys that allow blind spectroscopy, such as Euclid (Euclid Collaboration et al. 2025). The key technical innovation of this work lies in the use of gradient-boosting algorithms that can capture complex nonlinear relations in the data (e.g. Caruana & Niculescu-Mizil 2006). This makes them well suited for the task of preselecting LAEs.

This paper is organized as follows. Section 2 describes the data we used. Section 3 outlines the algorithmic framework and details the methods with which we evaluated the model, divided the data, and optimized the hyperparameters. In Section 4 we present the results of the machine-learning models. This includes an in-depth analysis of the training performance, predictions, and robustness to observational uncertainty perturbations. In Section 5 we discuss the main results and also discuss the potential applications of this work to other surveys. Finally, Section 6 offers a summary of the key findings.

We adopt the same ΛCDM cosmological model as Weaver et al. (2022) and Sobral et al. (2018b) ($H_0 = 70$ kms^{-1} Mpc^{-1}, $\Omega_M = 0.3$, and $\Omega_\Lambda = 0.7$). The magnitudes are given in the AB system (Oke 1974).

2. Data

We required a statistically significant sample of sources from LAE and non-LAE (nLAE) populations, where the latter represents other typical SFGs. They are both necessary for training a classifier to distinguish between them based on their broadband photometric properties. To achieve this, we used the SC4K (Sobral et al. 2018b) sample together with the COSMOS2020 catalog (Weaver et al. 2022).

2.1. SC4K

SC4K (Sobral et al. 2018b) is a survey in the COSMOS field (Capak et al. 2007; Scoville et al. 2007) that is designed to identify LAEs at $z \in [2, 6]$. It uses 12 medium and 4 NBs over a 2 deg^2 area, covering a comoving volume of $\sim 10^8$ Mpc^3. This sample includes 3,908 sources that were selected using an observed EW threshold of EW $> 50 \times (1 + z)$Å and $\Sigma > 3$, where the latter is the emission-line or excess significance (e.g. Bunker et al. 1995), which measures the degree to which the observed counts in a broadband filter deviate from the expected counts based on measurements in a NB filter. The typical EW threshold is 25Å (e.g. Santos et al. 2016). With almost twice this value, contamination by lower redshift line-emitters is less likely. We also grouped these sources (according to the corresponding medium-band filter) into specific redshifts following Table 3 in Sobral et al. (2018b). For each selection filter, we attributed the average value of the redshift interval (see Sobral et al. (2018b) for a detailed description of the sample selection.)

2.2. COSMOS2020

In addition to LAEs, we required a sample of nLAE sources (i.e., sources that are present in COSMOS2020 but not in SC4K) within the same field and redshift range to ensure a well-balanced dataset for comparative studies. For this purpose, we used the COSMOS2020 CLASSIC catalog (Weaver et al. 2022), which is based on the COSMOS survey (Scoville et al. 2007). We did not use the FARMER catalog because we required reliable photometry, which is primarily restricted to the UltraVISTA (McCracken et al. 2012) footprint as a result of mask requirements. Although it has fewer sources, CLASSIC ensures a complete and uniform spatial coverage for the analysis (i.e., for a comparison with the SC4K; see Weaver et al. (2022)).

We started by restricting the photometric redshift range ($2 < z < 6$) of the COSMOS2020 sample to match the SC4K survey because we lack comparable LAE samples outside this range to construct a meaningful classifier. The colors and intrinsic fluxes are highly dependent on redshift. We also excluded the SC4K sources from the COSMOS2020 sample. We cross-matched the two catalogs using TOPCAT (Taylor 2017) (with a maximum matching error of 1") to integrate photometric information into SC4K. Following these steps, COSMOS2020 and SC4K (hereafter LAE sample) contained 196,713 and 3,346 sources, respectively.

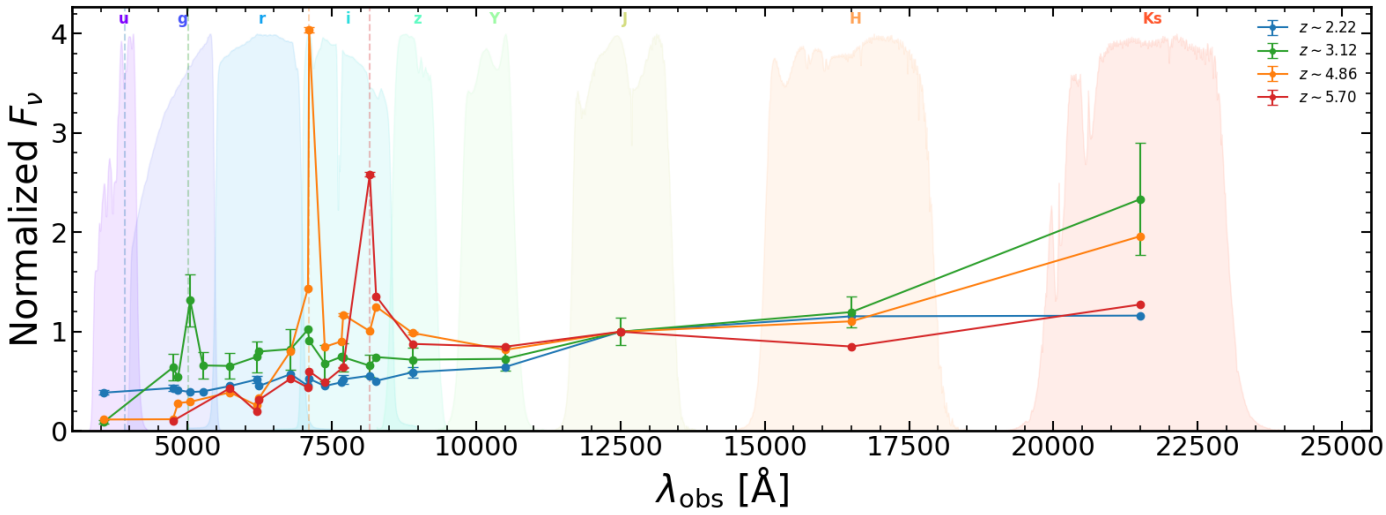


Fig. 1: Median SEDs of LAEs at different redshifts and normalized filter profiles of the broad bands. We also include the broadband transmission curve as background shading for context, and the vertical line in the NB central wavelength shows the expected location of Lyα.

We assumed that galaxies in the COSMOS2020 catalog at 2<z<6 that were not detected in SC4K are not LAEs. We expected some LAE contaminants from COSMOS2020, however, in particular, sources with Lyα emission with low EWs, because SC4K relies on human validation, which is prone to failure. It also only covers certain redshift slices, possibly missing LAEs outside of these intervals. To mitigate this issue, and because LAEs are a subdominant population (e.g. Cassata et al. 2015), we selected five different subsets and did not rely on a single sample. This was to ensure that first, the number of nLAEs extracted from COSMOS2020 was equal to the size of the LAE sample to exclude imbalances in the sample that might bias the model toward the majority class (e.g. Phelps et al. 2025). We also ensured that nLAE sources were drawn from the COSMOS2020 sample, following the same i-band magnitude and redshift distribution as the LAE sources. Specifically, the i-band magnitude ranges from 18 to 34 (minimum and maximum values for the i-band magnitude in the LAE sample, respectively), divided into intervals of one magnitude each. The redshift ranges from 2 to 6, also divided into intervals of one unit. This ensured that any potential differences are not dominated by distributions in redshift or UV brightness from our choice of samples. Following these steps, we obtained five subsamples that each contained the same LAE sources and different nLAE sources, extracted as described above.

The magnitude distributions in the two samples show (Figure A.1) that the LAE sample in the g band contains brighter sources than the nLAE sample. This highlights the possible bias toward brighter LAEs, which stems from that fact that the LAE sample was obtained using NB filters. On the other hand, no significant differences were found between the populations in the other bands. With the exception of the g, r, i, and z bands, there is a significant number of missing values (e.g., NaN) in both samples, which can either signify nondetections or no image coverage. A nondetection is often essential for selecting high-redshift objects (e.g., Lyman-break selections). Moreover, the gradient boosting algorithms we used can effectively handle missing values, and these consequently pose no problem for our work.

3. Machine-learning framework

In light of the accelerated growth in size and complexity of astronomical datasets, astronomers are developing automated tools to detect, characterize, and classify objects using these rich and complex datasets (e.g. Baron 2019; Fluke & Jacobs 2020; Carvajal et al. 2023; Humphrey et al. 2023). Our focus here is supervised learning to train models that can accurately classify LAEs based on broadband photometric labels.

3.1. Feature engineering

The selected features for our work consist of broadband photometric properties: fluxes, magnitudes, and colors in the optical and near-infrared (NIR) spectra. The nine bands and their corresponding instruments, central wavelength, and observed depth are listed in Table 1. For each band, we directly extracted fluxes and magnitudes with aperture magnitude APER3 from COSMOS2020, and we calculated the unique permutations of the broadband colors using the magnitudes. This resulted in 54 features. We worked with the optical and NIR properties because they help us in our overarching goal of creating a model that is applicable to other large surveys that are limited to these wavelengths.

The data-splitting strategy we followed involved dividing the data into 75% for the training set, 10% for the validation, and 15% for the testing (following the division suggested in Joseph 2019). This also ensured a stratified and shuffled division.

Figure 1 shows the median spectral energy distributions (SEDs) of LAEs at different redshifts, normalized to the J band, overlaid with the normalized transmission curves of the broadband filters we used. For each NB redshift bin in COSMOS2020 (NB392, NB501, NB711, and NB816), we selected galaxies with at least six photometric bands with positive flux. For each filter, the median flux across these galaxies was computed based on the positive flux values, and the uncertainty was estimated using the median absolute deviation divided by the square root of the number of galaxies that contribute to that filter (NB392: 100 valid sources, NB501: 11 valid sources, NB711: 42 valid sources, and NB816: 108 valid sources). The median SED was then normalised by the flux in a reference band redward of Lyα

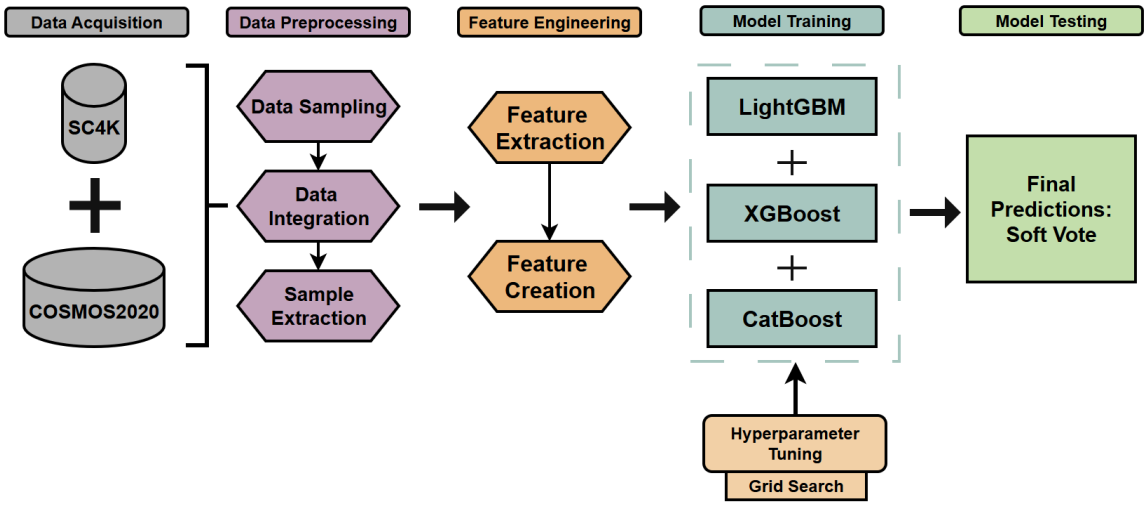


Fig. 2: ML flowchart of our framework. It begins with the acquisition of the publicly available SC4K and COSMOS2020 catalogs. The acquired data are then preprocessed, which involves sampling, integrating, and extracting the samples to be used by the ML models. This is followed by extraction of fluxes and magnitudes as features together with the creation of colors using the latter. The training phase involves the optimization of hyperparameters using a grid search for each algorithm: LightGBM, XGBoost, and CatBoost. Finally, the final target label is obtained by combining all the individual predictions into a single prediction using soft voting.

Table 1: Photometric bands and the corresponding instrument, central wavelength, and observed depth (3″).

Band	Instrument/Telescope	λ [Å]	Depth [mag]
u^*	MegaCam/CFHT	3858	27.1
g	HSC/Subaru	4847	27.5
r	HSC/Subaru	6219	27.2
i	HSC/Subaru	7699	27.0
z	HSC/Subaru	8894	26.6
Y	VIRCAM/VISTA	10216	26.1
J	VIRCAM/VISTA	12525	25.9
H	VIRCAM/VISTA	16466	25.5
K	VIRCAM/VISTA	21557	25.2

Notes. The instruments we used are the Canada-France-Hawaii telescope (CFHT) and MegaCam instrument (Boulade et al. 1998), the Hyper Suprime-Cam (HSC) (Miyazaki et al. 2018), and the VIRCAM instrument on the VISTA telescope from the UltraVISTA survey (McCracken et al. 2012).

(J band) for a direct comparison of the shape and emission excess (i.e., Lyα bump) in the bins.

3.2. Algorithms

We used ensemble learning (e.g. Ganaie et al. 2022; Cunha & Humphrey 2022; Euclid Collaboration et al. 2023; Cunha et al. 2024), which is an approach in machine learning that seeks a better predictive performance by combining the outputs of several models. In contrast to ordinary learning approaches that try to construct one learner from training data, ensemble uses a set of models and then aggregates their predictions in order to achieve a stronger predictive model with a better accuracy and an increased robustness (e.g. Sagi & Rokach 2018). Within ensemble learning, we used gradient-boosting algorithms (Friedman 2000; Natekin & Knoll 2013), where the base learners are generated sequentially and thus explore the dependence between the base learners. This dependence can be used to good advantage as the performance can be boosted by minimizing the errors made by the previous model. This improves the overall performance.

We implemented three widely used current (Florek & Zagdański 2023) gradient-boosting algorithms: CatBoost (version 1.2.3; Prokhorenkova et al. 2018), XGBoost (version 4.3.0; Chen & Guestrin 2016), and LightGBM (4.3.0; Ke et al. 2017). These models have been widely used for industry and research problems (e.g. Parsa et al. 2019; Huang et al. 2019; Ponsam et al. 2021). They all belong to the family of gradient-boosting decision trees, which build an ensemble of decision trees in a sequential manner. Each new tree is trained to correct the prediction errors made by the ensemble so far, using gradients of the loss function (e.g. Natekin & Knoll 2013). The key idea behind boosting is that while each individual tree may be a weak learner, their collective combination results in a strong and flexible model.

Despite this shared foundation, the three frameworks differ in several key structural aspects. CatBoost creates symmetric trees with the same splitting rules at each level, which helps with speed and reduces overfitting. In contrast, LightGBM and XGBoost build asymmetric trees that can have different split conditions at the same depth. When it grows trees, LightGBM uses a leaf-wise method that focuses on the branches with the highest potential to reduce error. This makes the models smaller and faster, but increases the risk of overfitting. CatBoost and XGBoost use a level-wise approach and grow all branches evenly. Regarding splitting methods, CatBoost checks all feature-split options for each leaf and selects the option with the least penalty. LightGBM speeds the computation up by focusing on data points with the largest errors. XGBoost uses a slower histogram-based method to determine the best splits. These differences mean that a combination of the three algorithms through soft voting takes advantage of their complementary strengths and helps us to reduce individual model biases.

To determine the best combination of the hyperparameters, we used Grid Search (Yu & Zhu 2020), which systematically explores a predefined set of values and selects the combination that

optimizes a chosen performance metric. Not all the hyperparameters have the same effect on the performance variation of the algorithms (e.g. Bentéjac et al. 2019; Sipper 2022). We therefore followed the suggestions from the documentation of the three algorithms[1] for guidance about the most effective hyperparameters for each. All the hyperparameters we optimized are listed in Appendix B. The summary of the optimization procedure is shown in Table B.1, and the default hyperparameters and grids we used are described for each algorithm. Table B.2 lists the final set of hyperparameters that was obtained using Grid Search for each algorithm and sample.

Figure 2 shows a flow diagram that contains an overview of the whole framework.

3.3. Evaluation metrics

We wished to evaluate the predictive ability, generalization capability, and overall quality of our models. In this subsection, we briefly introduce the evaluation metrics that we used mainly. The scores for the metrics we present below varied between zero and one, where one is the maximum score.

Precision (also called purity) is the fraction of relevant instances (known as true positives) among the retrieved instances,

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{1}$$

where TP is the number of true positives in a given class, and FP is the number for false positives in a given class. Recall, also known as completeness, is the fraction of relevant instances that were retrieved,

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{2}$$

where FN is the number of false negatives in a given class.

These two metrics are different, and it is therefore important to understand they are correctly interpreted and which importance should be assigned to each. Precision is more important when the cost of false positives is high, that is, in situations when misclassifying an instance as positive has serious consequences. On the other hand, recall is more effective when the cost of false negatives is high, that is, when the consequence of misclassifying an instance as negative is relevant. Because they are different and their associated meaning differs, it is useful to combine them using the F1-score, which is the harmonic mean of precision and recall,

$$\text{F1} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

The objective is to identify a substantial number of LAEs while minimizing false positives. The F1-score is therefore beneficial because it provides a balance between these two factors. Accuracy is another objective. This is the fraction of predictions that the model derived correctly. It is formally defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}, \tag{4}$$

and it should be used preferably when the number of objects in each class is roughly the same. This is true for our samples.

We calculated these four metrics for the models we obtained because each one of them attributes different importances to TP, TN, FP, and FN.

[1] LightGBM, XGBoost, CatBoost

4. Results

In each one of the five samples presented in Section 2 (containing the same number of LAEs and nLAEs), the three algorithms were trained and tested. This resulted in 15 different models. In the following section, we present the performance evaluation of the algorithms with and without hyperparameter optimization together with the application of these algorithms to the remainder of the COSMOS2020 sample at $z \in [2, 6]$.

4.1. Machine-learning training performance and predictions in COSMOS2020

Each algorithm was initially trained and evaluated using default hyperparameters with various input features: fluxes only, both fluxes and magnitudes, and a combination of fluxes, magnitudes, and colors. When we only used fluxes or fluxes and magnitudes, the averaged F1-scores (for five samples) ranged from $\sim 0.82 - 0.84$ in the test, depending on the algorithm. The inclusion of colors significantly enhanced the model performances and resulted in higher F1-scores of $\sim 0.86 - 0.88$. We found a clear advantage in using a more extensive set of features to improve the predictive accuracy of the models, even though the physical information in the fluxes and magnitudes was essentially the same. This result is consistent with the findings by Cunha & Humphrey (2022); Euclid Collaboration et al. (2023); Cunha et al. (2024). The results we present below were obtained with fluxes, magnitudes, and colors as features.

The classification evaluation metrics we computed for each model without the hyperparameter optimization are presented in Table 2 by averaging the results of the five samples. CatBoost performed better than the other algorithms overall. The difference is not highly significant compared with LightGBM, but XGBoost slightly underperformed in comparison. The results of the three algorithms all have a lower recall ($\sim 85 - 86\%$) than the other three metrics. On the other hand, precision is clearly the highlighted metric ($\sim 88 - 89\%$).

We optimized the hyperparameters (described in Appendix B) for all models using LightGBM, XGBoost, and CatBoost. The results are summarized in Table 3. The validation set improved by $\sim 2 - 5\%$ when the hyperparameters were optimized compared to the default case. These improvements became marginal for the test set. In contrast to the default case, LightGBM slightly outperformed the others, especially in the test set. CatBoost performed more poorly in the test set than in the default case. In Appendix C we present more detailed results regarding the impact of redshift of the training and testing samples in the model performance. For this purpose, we show in Figure C.1 the average confusion matrices for each algorithm separated by redshift interval bins. The three algorithms tend to increase at $z > 4$ in general for the number of false predictions. This effect is present in the testing and training samples, but the latter is less affected.

The models were applied to the COSMOS2020 dataset with a focus on sources within $z \in [2, 6]$ while excluding the five samples we used for the training. This resulted in 179,469 previously unused sources (hereafter called the predictions sample). We evaluated how often these sources were predicted as LAEs in 15 models. Over 70% of sources were not classified as LAEs by any model (52,707 were predicted by at least one model), while fewer than 5% (7,073 sources) were consistently predicted as LAEs by all 15 models. The average prediction scores (ranging from zero to one) in the 15 models was also analyzed by computing the mean of the scores assigned by each model. By default, a

Table 2: Average validation and test-set classification evaluation metrics for LightGBM, CatBoost, and XGBoost without hyperparameter optimization, that is, with the default settings. The standard deviation of each metric is also included.

	Validation				Test			
Model	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec
LightGBM	0.859± 0.011	0.860± 0.012	0.869± 0.021	0.848± 0.010	0.871± 0.009	0.873± 0.008	0.887± 0.012	0.855± 0.012
XGBoost	0.859± 0.016	0.860± 0.018	0.870± 0.032	0.848± 0.009	0.865± 0.008	0.867± 0.008	0.879± 0.010	0.852± 0.009
CatBoost	0.864± 0.008	0.865± 0.009	0.872± 0.017	0.857± 0.004	0.873± 0.006	0.876± 0.006	0.890± 0.008	0.857± 0.005

Table 3: Average validation and test-set classification evaluation metrics for LightGBM, CatBoost, and XGBoost with hyperparameter optimization. The standard deviation of each metric is also included.

	Validation				Test			
Model	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec
LightGBM	0.875± 0.012	0.876± 0.013	0.888± 0.024	0.862± 0.006	0.870± 0.012	0.872± 0.012	0.889± 0.015	0.851± 0.011
XGBoost	0.871± 0.011	0.873± 0.013	0.886± 0.024	0.858± 0.005	0.866± 0.011	0.869± 0.011	0.885± 0.014	0.845± 0.011
CatBoost	0.866± 0.015	0.868± 0.013	0.878± 0.015	0.856± 0.030	0.863± 0.012	0.866± 0.010	0.882± 0.008	0.845± 0.022

source was classified as an LAE when its mean score exceeded 0.5. Scores closer to zero corresponded to nLAEs. We verified that the prediction score of 21,440 sources was higher than 0.5, within which the score for 3627 LAE candidates was higher than 0.9.

4.2. Validating predicted LAEs with spectroscopic information

After predicting LAEs in the COSMOS field, it is relevant to validate some of these predictions using publicly available spectroscopic catalogs. This validation provides insights into the decision-making process of the model and helps us to confirm the nature of certain sources. For this purpose, we used the COSMOS spectroscopic redshift compilation[2] (Khostovan et al. 2025). It contains redshifts for 97,929 unique objects from 108 different observing programs up to $z \sim 8$.

We cross-matched the predicted LAEs from our model with the entries in this compilation using TOPCAT (Taylor 2017), which applies the SKY algorithm with a maximum matching error of 1″. We restricted the results for $z \in [2, 6]$ and applied a quality flag $Q_f > 1$, which ensured a confidence level $\geq 80\%$. We also excluded all surveys without public information or without wavelength coverage for a *Ly*α detection. A source was considered a match when it had a rounded average prediction score higher than 0.5 and was confirmed in any of the surveys available in Khostovan et al. (2025). Following these steps, we found 7 LAE confirmations in the VIMOS Ultra-Deep Survey (VUDS) (Le Fèvre et al. 2015; Tasca et al. 2017), 15 in the COSMOS Lyα Mapping And Tomography Observations (CLAMATO) Survey (Lee et al. 2018; Horowitz et al. 2022), and 8 in the Deep Imaging Multi-Object Spectrograph (DEIMOS) (Hasinger et al. 2018) with a Lyα EW higher than 20Å. Additionally, 10 were found in Schmidt et al. (2021), all identified as Lyα emitters by the authors and with a Lyα EW higher than 20Å. In the Hobby-Eberly Telescope Dark Energy Experiment (HETDEX) (Gebhardt et al. 2021; The HETDEX collaboration et al. 2023; Davis et al. 2023), 15 were identified as LAEs. Finally, 3 sources were studied in Rosani et al. (2020), 1 by Ning et al. (2020), and 1 source was presented in Pentericci et al. (2018). In total, we obtained the cross-match of 60 LAEs with public spectroscopic catalogs in the COSMOS field. This corroborates that our models are able to effectively identify LAE candidates. The list of all the sources is provided in Appendix D, and three example spectra are shown in Figure D.1.

[2] COSMOS Spectroscopic Redshift Compilation

4.3. Assessing the robustness of the classification models

Instrument measurements always include a degree of uncertainty over the measured magnitudes/fluxes. We assessed the impact of the level of uncertainty in our models by retraining the model with a dataset in which the magnitudes/fluxes were perturbed by their observational uncertainties. In each training/test sample, every source was perturbed by creating a Gaussian distribution with a mean equal to the flux value and a sigma equal to the flux error. We then drew a random value out of this distribution. The perturbed fluxes were then converted into AB magnitudes, which in turn were used to calculate the colors. Previously, fluxes and magnitudes were extracted from COSMOS2020, but we now wished to propagate the pertubations from the fluxes into the magnitudes. This justifies the conversion. This was done for every source in the five samples 100 times each for every band. We thus created five perturbed samples that were then used to train and test new models that were then compared with the unperturbed models we presented and discussed above.

The perturbations can be separated into three different cases:

1. Perturbations in the input data. The new models were trained and tested in these new perturbed samples. We therefore ought to investigate the effect of the perturbations on the model performance and the consequences when predicting in unseen data.
2. Perturbations in the predictions table. We also introduced these variations in the predictions sample and then applied the unperturbed models to these 100 perturbed predictions tables.
3. Joint perturbations. We joined the perturbations performed in steps 1 and 2, that is, we applied the perturbed models to the perturbed predictions tables.

We found a decrease in the F1-score by $\sim 5\%$ when we introduced perturbations compared to the unperturbed case, but no variation between iterations. This decrease was expected as a result of the introduction of variance in the data, but the F1-score still remained above 0.81 for all algorithms. We also investigated

the variation in the number of predicted LAEs with a prediction score over 0.9 in the perturbed models for the three different cases. The number of predicted LAEs for the input data (1) decreased by $\sim$ 20%, while the perturbations in the prediction data (2) led to an increase of $\sim$ 40% in the number of sources with a score higher than 0.9. When the two perturbations were joined (3), the number of predicted LAEs was similar to the unperturbed case. The significant increase in the number of predicted LAEs when the prediction data were perturbed indicates that there might be a shift toward higher values of prediction scores in this case. This can also be affect by the fact that the prediction table is larger by almost two orders of magnitude than the input data. This leads to a higher chance of generating a LAE classification.

In order to further investigate this possible shift, we show in Figure 3 two-dimensional histograms of the perturbed and unperturbed prediction scores that we computed for the three perturbation cases. For the first perturbation case (1), the data points seem to follow a 1:1 relation with some scatter around this line, but there is some trend toward lower perturbed scores for sources with high unperturbed scores. The results clearly start to deviate from the 1:1 relation when the predictions table is perturbed (2), where in addition to the trend at the lower right side of the plot, there is also a clear shift of sources with low unperturbed scores toward higher perturbed scores. In the last case, the excess diminishes, but the trend at the lower right side of the plot is stronger.

To determine the cause of the stronger variation in the prediction data perturbations, we repeated the perturbations in the prediction data and separated them into three cases, depending on the feature set: (i) fluxes, (ii) fluxes and magnitudes, and (iii) fluxes, magnitudes, and colors. The results are compiled in Figure 4, separated for each case. When the fluxes alone were considered, the 0.25 range contains 237 sources. When magnitudes were added to the feature set, this number increases to 763. Finally, for the fluxes, magnitudes, and colors, the number of outliers increases massively to 5545. As the number of features increases, the scatter around the 1:1 relation between unperturbed and perturbed prediction scores therefore increases as well, which means that the bumps in the left and right parts of the plots become more pronounced. This is especially true when colors are introduced, which might be related to the double perturbation needed to compute each color. Colors imply two independent and random perturbations, one per band, whereas for fluxes/magnitudes, only one error/perturbation was considered.

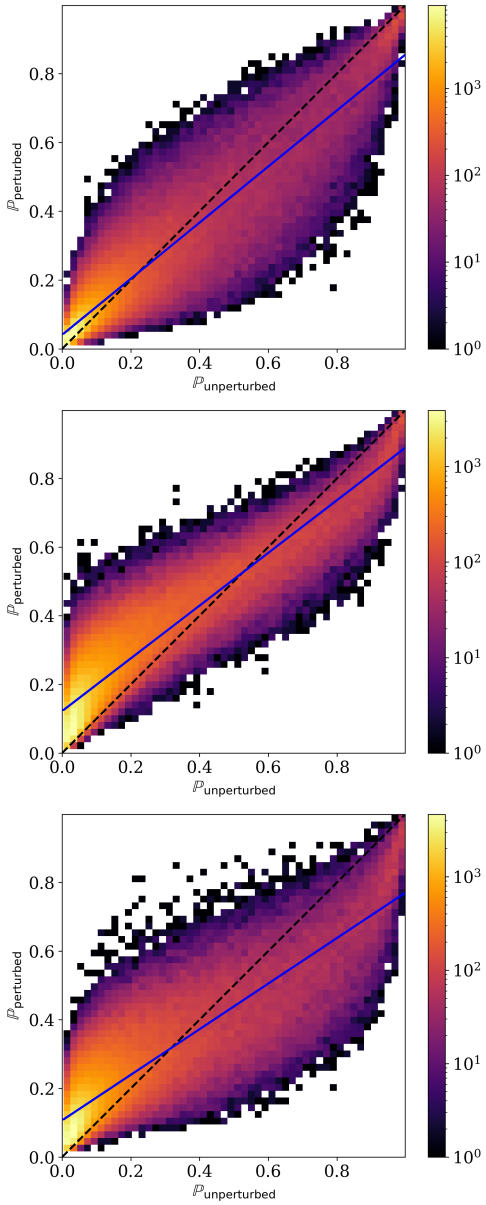
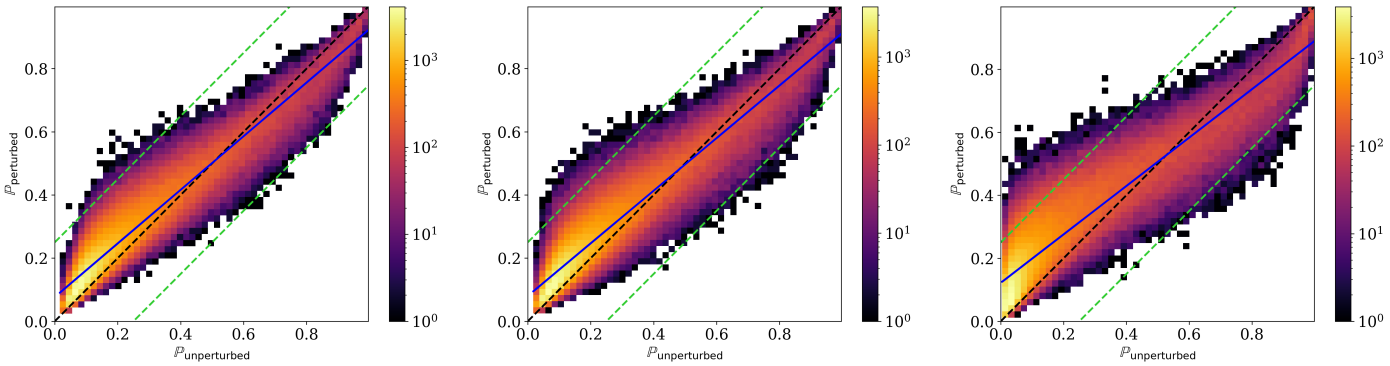


Fig. 3: Comparison of the perturbed prediction scores and unperturbed scores for the input data perturbation, prediction table perturbation, and joint perturbation cases in the panels from top to bottom, respectively. Average prediction scores over all the models were computed for each source. The blue lines represent a linear fit to the data points, and the dashed black lines show the 1:1 relation.

5. Discussion

The metrics for the default (Table 2) and optimized models (Table 3) show robust models that are able to identify LAE candidates based on broadband photometry data in the optical and NIR alone. The small improvements with hyperparameter optimization are consistent with reports inother works (e.g. Probst et al. 2018; Bahmani et al. 2021), which might indicate that the default algorithms are already near the maximum performance for these algorithms and these data.

More than 50,000 sources were predicted as LAE candidates in COSMOS2020 by at least one model, and approximately 7,000 of these were predicted by all 15 models we trained. Even when we only consider the latter, this is a highly significant number of predicted sources. If it were confirmed, it might more than double the existing SC4K sample. This confirms that previously unknown LAEs candidates might be detected in this way. We spectroscopically validated 60 LAEs in the COSMOS field (Section 4.2), which again highlights that the models fulfill their potential in complementing traditional LAE detection methods. A possible explanation for the failure to detect these sources using NB and intermediate-band (IB) surveys might be that NB/IB filters only cover limited redshift ranges. They thus risk losing many sources between the gaps left at $z \in [2, 6]$, which can now be explored (see Paulino-Afonso et al. in prep) because the whole redshift range can now be searched with the help of these models.



(a) Fluxes (b) Fluxes and magnitudes (c) Fluxes, magnitudes, and colors

Fig. 4: Comparison of the perturbations in the prediction data results for fluxes (a), fluxes and magnitudes (b), and fluxes, magnitudes, and colors (c). The blue lines in each subplot represent a fit to the data, the dashed black lines show the 1:1 relation, and the dashed green lines show the 0.25 score deviation from the black line.

5.1. Effect of the redshift on the model performance

Figure 5 presents the test and train F1 scores as a function of redshift for the three algorithms. The figure shows that all three models performed strongly in the lower redshift bins (approximately $z \in [2, 4]$), with F1 scores ranging from approximately 0.9 to 1.0. With increasing redshift ($z > 4$), however, the F1 scores consistently decrease in all models. This indicates a general degradation in the predictive performance. The overplotted histogram further confirms that the distribution of training and testing data is strongly skewed toward low-redshift sources, with a marked drop in the sample availability beyond $z \sim 4$.

This relation underscores a common challenge in machine learning: performance degradation in underrepresented regimes. The limited availability of high-redshift sources hampers the model performance when the data are scarce. The nearly parallel decline of the F1-scores in all three models suggests that architectural or optimization differences alone are insufficient to counteract the effects of this data imbalance.

To mitigate the performance drop at high redshifts, several strategies might be considered in principle. First, physically motivated data augmentation using the spectral synthesis code CLOUDY (Chatzikos et al. 2023) might be employed, for example. This method enables the generation of synthetic spectra based on astrophysical parameters, which can be redshifted and transformed into model-compatible features. While this approach introduces valuable data where observations are limited, it is limited, including potential mismatch with real observational noise, reliance on uncertain priors, and high computational cost. Nonetheless, when it is carefully applied, it might improve the model performance in data-sparse regimes.

An alternative to addressing the redshift imbalance is reweighting the training samples based on the inverse frequency of their redshift bin. This gives more weight to underrepresented high-redshift data and would improve the model performance in sparse regions without creating synthetic data. This method assumes that all redshift bins are equally informative and can cause overfitting or instability if high-weighted samples are noisy, however,. Additionally, a weight estimation can be unreliable in high-dimensional or limited data settings. Despite these challenges, reweighting can enhance the robustness across redshift intervals.

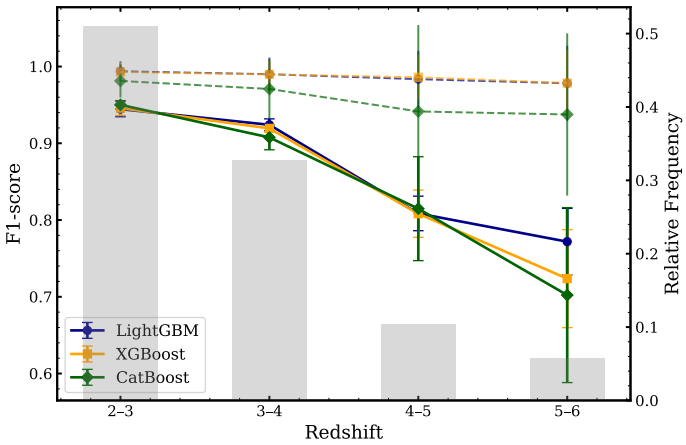


Fig. 5: Variation in the test and train (dashed) F1-scores as a function of redshift for the three gradient-boosting models LightGBM, XGBoost, and CatBoost. The overlaid histogram represents the normalized relative frequency of sources within each redshift bin in the training/testing samples. This combined view illustrates the relation between the model performance and data availability.

5.2. Potential applicability in the era of large surveys

One important goal of this work was to explore whether the models might be applied to other fields and surveys. This generalization is not straightforward as it involves conversion between different sets of filters, with different transmission curves. Our oversimplified approach aims to evaluate how these models perform when certain bands that we used are unavailable. As mentioned, our study used filters that observed COSMOS and covered wavelengths from 3560 Å to 23094 Å. Accordingly, we removed photometric bands from the feature set given to train the models and tried to mimic the filter wavelength coverage of potentially interesting surveys and missions to which this work might be applied. We provide a summary of each survey we considered for this exercise and the filter set we used to train the models in each case below.

– Euclid. The Euclid mission (Euclid Collaboration et al. 2025) will observe 15,000 deg^2 of extragalactic sky from the Sun–Earth Lagrange point L2. Among many other

Table 4: Average evaluation metrics in the test set compared with using a filter set adapted according to the wavelength coverage available in other surveys: Euclid, LSST, DES, and Roman. The standard deviations for each metric are also included.

	Test set				
Survey	Algorithm	F1-Score	Accuracy	Precision	Recall
All filters	LightGBM	0.870±0.012	0.872±0.012	0.889±0.015	0.851±0.011
	XGBoost	0.866±0.011	0.869±0.011	0.885±0.014	0.845±0.011
	CatBoost	0.863±0.012	0.866±0.010	0.882±0.008	0.845±0.022
Euclid	LightGBM	0.714±0.014	0.722±0.013	0.735±0.014	0.693±0.016
	XGBoost	0.714±0.016	0.724±0.019	0.740±0.025	0.690±0.012
	CatBoost	0.719±0.010	0.728±0.015	0.745±0.026	0.694±0.005
LSST	LightGBM	0.864±0.008	0.867±0.008	0.887±0.009	0.842±0.010
	XGBoost	0.866±0.006	0.869±0.006	0.890±0.010	0.842±0.010
	CatBoost	0.861±0.008	0.865±0.006	0.887±0.008	0.838±0.018
DES	LightGBM	0.797±0.013	0.802±0.014	0.819±0.021	0.777±0.007
	XGBoost	0.797±0.009	0.802±0.011	0.821±0.018	0.774±0.004
	CatBoost	0.800±0.007	0.806±0.007	0.825±0.014	0.777±0.014
Roman	LightGBM	0.825±0.009	0.829±0.010	0.846±0.016	0.805±0.011
	XGBoost	0.827±0.008	0.831±0.008	0.847±0.010	0.809±0.010
	CatBoost	0.823±0.010	0.828±0.010	0.845±0.017	0.803±0.017

data, the Euclid plans include determining the photometry of at least one billion galaxies. To achieve this goal, the Euclid spacecraft has a visible-wavelength camera (the VISible instrument; VIS) and a near-infrared camera/spectrometer (the Near-Infrared Spectrometer and Photometer; NISP). The VIS instrument covers the wavelength between 5500–9000Å, and NISP has three broadband filters (Y, J, and H) that cover the wavelength ranges of 9000–11920Å, 11920–15440Å, and 15440–20000Å, respectively. To mimic the wavelength coverage of the VIS instrument, we considered the r, i, and z bands, and to mimic the coverage of the NISP instrument, we considered the Y, J, and H bands.

- LSST. The LSST (Ivezić et al. 2019) will cover about 18,000 deg^2 of the southern sky with six filters (u, g, r, i, z, and Y) in a wavelength interval from approximately 3500 to 10600Å. Consequently, only the homonymous bands available in the COSMOS2020 catalog were used, that is, the u*, g, r, i, z, and Y bands.
- DES. The Dark Energy Survey (DES) (Dark Energy Survey Collaboration et al. 2016) astronomical survey is designed to study hundreds of millions of galaxies and to help us to understand the nature of dark energy. It uses the newly built Dark Energy Camera (DECam), which records images using five filters (g, r, i, z, and Y) that span wavelengths from 4000Å to 10800Å. We therefore only used the g, r, i, z, and Y bands.
- Roman. The Nancy Grace Roman Space Telescope (Spergel et al. 2015) is a NASA infrared space telescope created to answer important questions in the areas of dark energy, exoplanets, and infrared astrophysics. It is equipped with a wide-field instrument (WFI) carrying eight filters spanning 4800 – 23000 Å. The covered wavelength range is large, and only the u* band was removed from our exercise.

After defining the changes in the features used to mimic the wavelength coverage of each survey presented above, we trained the models and ran them for each survey. The metrics we obtained for the test set are summarized in Table 4.

We compared the results for Euclid and Roman, where the only differences were the removal of the u^* and g bands in Euclid and only the u^* band in Roman. The effect from excluding the u^* band is significant, as is reflected in the Roman results ($\sim$ 5% decrease). This effect is even more pronounced in Euclid, where the additional removal of the g-band further amplified the difference significantly ($\sim$ 17%). These findings highlight the crucial role of these two bands in enhancing the performance of our models to accurately distinguish LAEs from non-LAEs. In contrast, the removal of the J, H, and K bands had little effect on the performance, as evidenced by the LSST results ($<$ 1%). This is also visible in DES results compared to Roman, where the difference is the inclusion of the J, H, and K bands in the latter. This gained $\sim$ 2% in the F1-Score value.

Our analysis emphasized the high importance of the g, and u* bands. When these bands were included, we were better able to measure the Lyman-break at $z \geq 3$. This probably indicates that coverage of the Lyman-break in the photometry is important to assess the nature of these galaxies. Surveys that lack these bands will not be able to ensure a comparable classification performance. The NIR bands are less effective at distinguishing LAEs than optical bands, however. The reason might be that the Y, J, H, K bands have a high fraction of missing data (Figure A.1), which reduces their contribution to the classification models. This means that the filter set is almost not affected at all when these bands are removed.

Through a feature importance analysis (Figure E.1) of each algorithm, averaged over the five samples, we found that the g-r color is the most important feature in all the three algorithms. This color is a rough approximation of the β slope (e.g. Bouwens et al. 2009) at $z \sim 2-3$ (the bulk of the samples), which traces the attenuation by dust in the galaxy. At higher reshifts (≥ 3.5), it starts to measure the Lyman break, which is an indirect measure of the redshift (c.f. Steidel et al. 1996) and the neutral gas content. If g-r is the most important feature, it likely means that dust is crucial for separating LAEs and nLAEs (Santos et al. 2020). This is consistent with dust being one of the primary materials that absorbs Lyα photons and prevents them from escaping from galaxies.

6. Conclusion

We applied machine-learning techniques, specifically, the gradient-boosting algorithms LightGBM, XGBoost, and CatBoost, to identify LAEs candidates using broadband photometric data (fluxes, magnitudes, and colors) in the optical and NIR. Using SC4K and COSMOS2020, we extracted five samples with similar redshift and i-band magnitude distributions to ensure that we had comparable LAE and nLAE populations. We finally trained, tested, and analyzed the three algorithms in each one of the five samples, resulting in 15 models. We summarize our main conclusions below.

- The machine-learning models we developed demonstrated robust classification abilities with F1-scores from approximately 86% to 88% in the different algorithms. We also verified that our models presented a significant performance decrease at $z > 4$, where training/testing data are scarce.
- These models successfully identified over 7,000 LAEs candidates in unanimous agreement in all models, and over 50,000 were predicted by at least one model. This significantly increased the number of LAEs in the COSMOS field.
- Validation with spectroscopic data from Khostovan et al. (2025) confirmed 60 of our machine-learning predicted LAEs. This underscored the effectiveness of the approach.
- The robustness of the models was tested against perturbations in data input by introducing the errors over the measured fluxes, which were then propagated to magnitudes and colors. This led to a small decrease of 5% in the F1-score and showed that with more features, the scatter between the perturbed and unperturbed predictions increased.
- We also found that removing the u^* and g bands highly affected the model performance (a decrease in the F1-score of $\sim 6-17\%$). This is problematic for surveys such as Euclid, LSST, and Roman. On the other hand, removing the most NIR bands (Y-,J-, and H-bands) had an effect of less than 1%. This shows that the LSST survey is well suited for the application of these models.

Throughout this work, we emphasized the capability of broadband photometric data, coupled with machine learning, to offer a cost-effective and scalable complement to traditional NB surveys and blind spectroscopy. This is extremely useful to process vast amounts of broadband photometric data quickly and to identify candidates over large sky areas. This is pivotal for upcoming data from large-scale surveys such as LSST and Euclid. Furthermore, we clearly understood that an immense number of additional LAEs can be identified beyond those already detected. This might expand the known population and contribute to a broader and more detailed map of the early Universe.

Data availability

All the data used throughout this work is publicly available[3]. The code needed to reproduce this work is also available on GitHub (`https://github.com/valeafonso/FLAEMING-Classification`), together with the ML models. The complete candidates table (with sources predicted by the 15 models) is only available in electronic form at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsweb.u-strasbg.fr/cgi-bin/qcat?J/A+A/.

Acknowledgements. We thank the anonymous referee for their valuable comments and suggestions, which have greatly improved this manuscript. This work was supported by Fundação para a Ciência e a Tecnologia (FCT) through the research grants UIDB/04434/2020 and UIDP/04434/2020, and the exploratory project EXPL/FIS-AST/1085/2021 (PI: Paulino-Afonso). APA acknowledges FCT support through the FCT Investigador FCT Contract No. 2020.03946.CEECIND. JF acknowledges FCT support through the Investigador FCT Contract No. 2020.02633.CEECIND/CP1631/CT0002. This work is 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 für Extraterrestrische 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. Based also on data obtained with the European Southern Observatory Very Large Telescope, Paranal, Chile, under Large Program 185.A-0791, and made available by the VUDS team at the CESAM data center, Laboratoire d'Astrophysique de Marseille, France. Some of the data presented herein were obtained at Keck Observatory, which is a private, non-profit organization operated as a scientific partnership among the California Institute of Technology, the University of California, and the National Aeronautics and Space Administration. The Observatory was made possible by the generous financial support of the W.M.Keck Foundation. This work was also based on the public SC4K sample of LAEs (Sobral et al. 2018b). The authors acknowledge the hard work of all members of each respective program used to validate spectroscopically the predictions. Finally, we are grateful for the publicly available programming language PYTHON, including the packages: Pandas (McKinney 2010), Numpy (Harris et al. 2020), Matplotlib (Hunter 2007), Seaborn (Waskom et al. 2020), and scikit-learn (Pedregosa et al. 2011).

[3] COSMOS2020, SC4K

References

Ajiki, M., Taniguchi, Y., Fujita, S. S., et al. 2003, AJ, 126, 2091
Bahmani, M., El Shawi, R., Potikyan, N., & Sakr, S. 2021, arXiv e-prints, arXiv:2108.13066
Baron, D. 2019, arXiv e-prints, arXiv:1904.07248
Bentéjac, C., Csörgo, A., & Martínez-Muñoz, G. 2019, Artificial Intelligence Review, 54, 1937
Boulade, O., Vigroux, L. G., Charlot, X., et al. 1998, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 3355, Optical Astronomical Instrumentation, ed. S. D'Odorico, 614–625
Bouwens, R. J., Illingworth, G. D., Franx, M., et al. 2009, ApJ, 705, 936
Bunker, A. J., Warren, S. J., Hewett, P. C., & Clements, D. L. 1995, MNRAS, 273, 513
Capak, P., Aussel, H., Ajiki, M., et al. 2007, ApJS, 172, 99
Caruana, R. & Niculescu-Mizil, A. 2006, in Proceedings of the 23rd International Conference on Machine Learning, ICML '06 (New York, NY, USA: Association for Computing Machinery), 161–168
Carvajal, R., Matute, I., Afonso, J., et al. 2023, A&A, 679, A101
Cassata, P., Tasca, L. A. M., Le Fèvre, O., et al. 2015, A&A, 573, A24
Chatzikos, M., Bianchi, S., Camilloni, F., et al. 2023, Rev. Mexicana Astron. Astrofis., 59, 327
Chen, T. & Guestrin, C. 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (New York, NY, USA: Association for Computing Machinery), 785–794
Ciardullo, R., Feldmeier, J. J., Krelove, K., Jacoby, G. H., & Gronwall, C. 2002, ApJ, 566, 784
Cunha, P. A. C. & Humphrey, A. 2022, A&A, 666, A87

Cunha, P. A. C., Humphrey, A., Brinchmann, J., et al. 2024, A&A, 687, A269
Dark Energy Survey Collaboration, Abbott, T., Abdalla, F. B., et al. 2016, MNRAS, 460, 1270
Davis, D., Gebhardt, K., Cooper, E. M., et al. 2023, ApJ, 946, 86
Dijkstra, M. 2014, PASA, 31, e040
Drake, A. B., Guiderdoni, B., Blaizot, J., et al. 2017, MNRAS, 471, 267
Dunlop, J. S. 2013, in Astrophysics and Space Science Library, Vol. 396, The First Galaxies, ed. T. Wiklind, B. Mobasher, & V. Bromm, 223
Euclid Collaboration, Humphrey, A., Bisigello, L., et al. 2023, A&A, 671, A99
Euclid Collaboration, Mellier, Y., Abdurro'uf, et al. 2025, A&A, 697, A1
Florek, P. & Zagdański, A. 2023, arXiv e-prints, arXiv:2305.17094
Fluke, C. J. & Jacobs, C. 2020, WIREs Data Mining and Knowledge Discovery, 10, e1349
Friedman, J. H. 2000, Annals of Statistics, 29, 1189
Fujita, S. S., Ajiki, M., Shioya, Y., et al. 2003, ApJ, 586, L115
Fynbo, J. P. U., Ledoux, C., Møller, P., Thomsen, B., & Burud, I. 2003, A&A, 407, 147
Ganaie, M., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. 2022, Engineering Applications of Artificial Intelligence, 115, 105151
Gawiser, E., Francke, H., Lai, K., et al. 2007, ApJ, 671, 278
Gebhardt, K., Mentuch Cooper, E., Ciardullo, R., et al. 2021, ApJ, 923, 217
Giavalisco, M., Ferguson, H. C., Koekemoer, A. M., et al. 2004, ApJ, 600, L93
Grove, L. F., Fynbo, J. P. U., Ledoux, C., et al. 2009, A&A, 497, 689
Hagen, A., Ciardullo, R., Gronwall, C., et al. 2014, ApJ, 786, 59
Harikane, Y., Ouchi, M., Shibuya, T., et al. 2018, ApJ, 859, 84
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357
Hasinger, G., Capak, P., Salvato, M., et al. 2018, ApJ, 858, 77
Horowitz, B., Lee, K.-G., Ata, M., et al. 2022, ApJS, 263, 27
Hu, E. M., Cowie, L. L., Barger, A. J., et al. 2010, ApJ, 725, 394
Huang, G., Wu, L., Ma, X., et al. 2019, Journal of Hydrology, 574, 1029
Huertas-Company, M. & Lanusse, F. 2023, PASA, 40, e001
Humphrey, A., Cunha, P. A. C., Paulino-Afonso, A., et al. 2023, MNRAS, 520, 305
Hunter, J. D. 2007, Computing in Science and Engg., 9, 90–95
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Joseph, V. R. 2019, Statistical Analysis and Data Mining: The ASA Data Science Journal, 15
Ke, G., Meng, Q., Finley, T., et al. 2017, in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17 (Red Hook, NY, USA: Curran Associates Inc.), 3149–3157
Khostovan, A. A., Kartaltepe, J. S., Salvato, M., et al. 2025, arXiv e-prints, arXiv:2503.00120
Kurk, J. D., Cimatti, A., di Serego Alighieri, S., et al. 2004, A&A, 422, L13
Law, D. R., Steidel, C. C., Shapley, A. E., et al. 2012, ApJ, 759, 29
Lawrence, A., Warren, S. J., Almaini, O., et al. 2007, MNRAS, 379, 1599
Le Fèvre, O., Tasca, L. A. M., Cassata, P., et al. 2015, A&A, 576, A79
Lee, K.-G., Krolewski, A., White, M., et al. 2018, ApJS, 237, 31
Matthee, J., Sobral, D., Santos, S., et al. 2015a, MNRAS, 451, 400
Matthee, J., Sobral, D., Santos, S., et al. 2015b, MNRAS, 451, 400
McCracken, H. J., Milvang-Jensen, B., Dunlop, J., et al. 2012, A&A, 544, A156
McKinney, W. 2010, in SciPy
Miyazaki, S., Oguri, M., Hamana, T., et al. 2018, PASJ, 70, S27
Nakajima, K., Ouchi, M., Shimasaku, K., et al. 2012, ApJ, 745, 12
Napolitano, L., Pentericci, L., Calabrò, A., et al. 2023, A&A, 677, A138
Natekin, A. & Knoll, A. 2013, Frontiers in Neurorobotics, 7
Ning, Y., Jiang, L., Zheng, Z.-Y., et al. 2020, ApJ, 903, 4
Oke, J. B. 1974, ApJS, 27, 21
Ono, Y., Itoh, R., Shibuya, T., et al. 2021, ApJ, 911, 78
Ouchi, M., Ono, Y., & Shibuya, T. 2020, ARA&A, 58, 617
Ouchi, M., Shimasaku, K., Furusawa, H., et al. 2003, ApJ, 582, 60
Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. K. 2019, Accident; analysis and prevention, 136, 105405
Partridge, R. B. & Peebles, P. J. E. 1967, ApJ, 147, 868
Paulino-Afonso, A., Sobral, D., Ribeiro, B., et al. 2018, MNRAS, 476, 5479
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825
Pentericci, L., Grazian, A., Fontana, A., et al. 2007, A&A, 471, 433
Pentericci, L., Vanzella, E., Castellano, M., et al. 2018, A&A, 619, A147
Phelps, N., Lizotte, D. J., & Woolford, D. G. 2025, arXiv e-prints, arXiv:2501.04903
Ponsam, J. G., Bella Gracia, S. J., Geetha, G., Karpaselvi, S., & Nimala, K. 2021, in 2021 4th International Conference on Computing and Communications Technologies (ICCCT), 634–641
Probst, P., Boulesteix, A., & Bischl, B. 2018, J. Mach. Learn. Res., 20, 53:1
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. 2018, in Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18 (Red Hook, NY, USA: Curran Associates Inc.), 6639–6649
Rosani, G., Caminha, G. B., Caputi, K. I., & Deshmukh, S. 2020, A&A, 633, A159
Runnholm, A., Hayes, M., Melinder, J., et al. 2020, ApJ, 892, 48
Sagi, O. & Rokach, L. 2018, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8
Santos, S., Sobral, D., & Matthee, J. 2016, MNRAS, 463, 1678
Santos, S., Sobral, D., Matthee, J., et al. 2020, MNRAS, 493, 141
Schmidt, K. B., Kerutt, J., Wisotzki, L., et al. 2021, A&A, 654, A80
Scoville, N., Aussel, H., Brusa, M., et al. 2007, ApJS, 172, 1
Shibuya, T., Ouchi, M., Harikane, Y., & Nakajima, K. 2019, ApJ, 871, 164
Sipper, M. 2022, arXiv e-prints, arXiv:2207.06028
Sobral, D., Matthee, J., Darvish, B., et al. 2018a, MNRAS, 477, 2817
Sobral, D., Santos, S., Matthee, J., et al. 2018b, MNRAS, 476, 4725
Spergel, D., Gehrels, N., Baltay, C., et al. 2015, arXiv e-prints, arXiv:1503.03757
Steidel, C. C., Giavalisco, M., Pettini, M., Dickinson, M., & Adelberger, K. L. 1996, ApJ, 462, L17
Taniguchi, Y., Shioya, Y., Ajiki, M., et al. 2003, Journal of Korean Astronomical Society, 36, 123
Tasca, L. A. M., Le Fèvre, O., Ribeiro, B., et al. 2017, A&A, 600, A110
Taylor, M. 2017, arXiv e-prints, arXiv:1707.02160
The HETDEX collaboration, Mentuch Cooper, E., Gebhardt, K., et al. 2023, Astrophysical Journal, 943, publisher Copyright: © 2023. The Author(s). Published by the American Astronomical Society.
Waskom, M., Botvinnik, O., Gelbart, M., et al. 2020, seaborn: Statistical data visualization, Astrophysics Source Code Library, record ascl:2012.015
Weaver, J. R., Kauffmann, O. B., Ilbert, O., et al. 2022, ApJS, 258, 11
Yu, T. & Zhu, H. 2020, ArXiv, abs/2003.05689

Appendix A: Magnitude distribution

In this appendix, we present the distribution of magnitudes, along with key statistics such as maximum, minimum, median, standard deviation, and the count of null values. This is presented separately for the LAE sample and the five nLAE subsamples in Figure A.1.

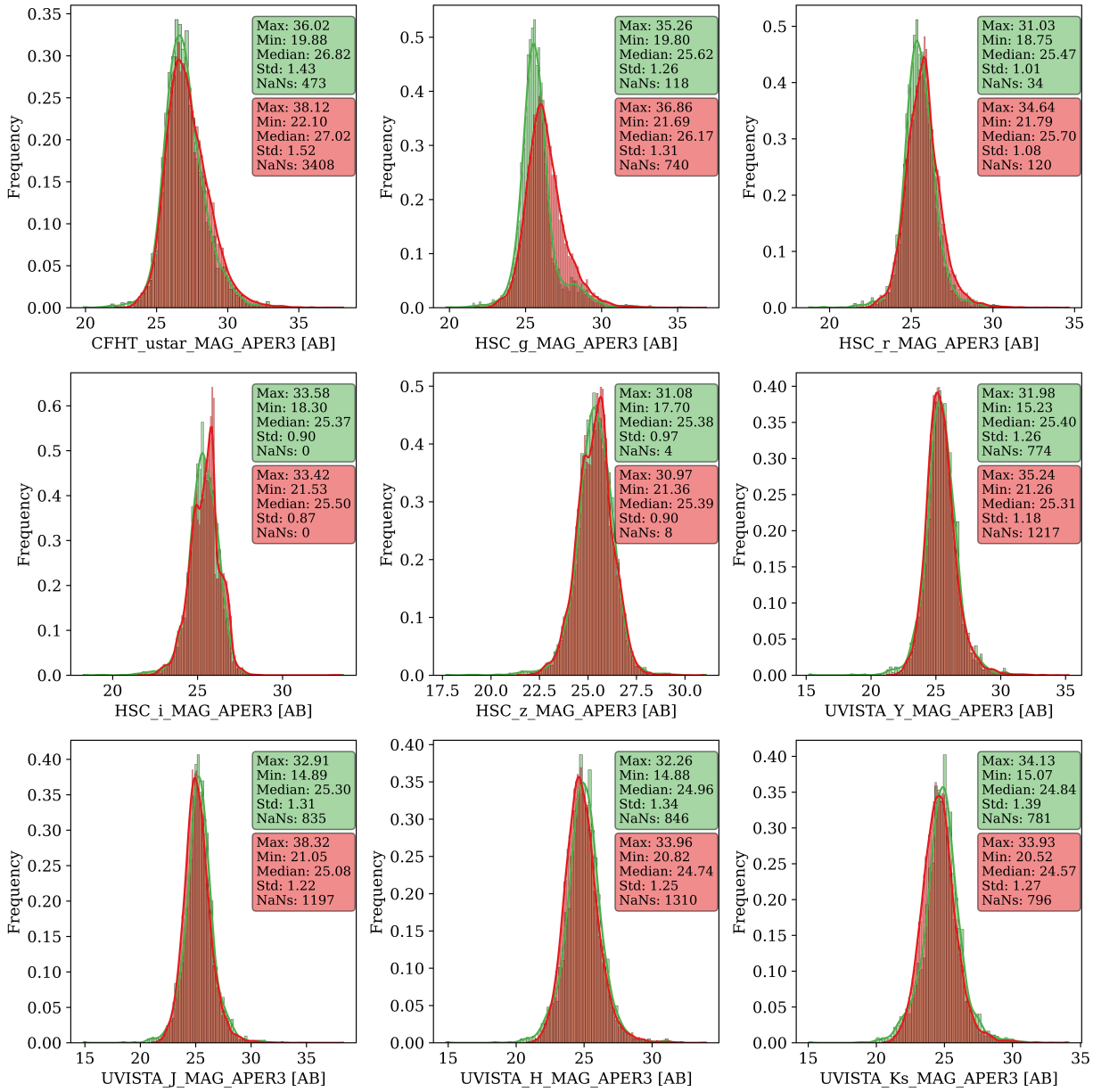


Fig. A.1: nLAE and LAE samples magnitudes distribution and statistics (maximum, minimum, median, standard deviation, and number of NaNs values). The LAE and nLAE samples are colored green and red, respectively.

Appendix B: Hyperparameter optimization grid

We present the grid used to perform the hyper-parameter optimization (Table B.1), together with a description of every hyper-parameter taken into consideration for optimization. In Table B.2 the best hyper-parameter set found for each algorithm and sample, ranked by F1-Score from GridSearch, is presented.

There are three hyper-parameters with a highlighted importance: *num_estimators* (number of boosting iterations), *learning_rate* (gradient step size), *depth* (maximum depth of each trained tree), and *subsample* (percentage of rows used per iteration to fit each tree).

Table B.1: Grids used in the hyper-parameter optimization procedure.

Algorithm	Default hyper-parameters	Optimization grid
LightGBM	n_estimators = 100 learning_rate = 0.1 max_depth = -1	n_estimators = {50,100,250,500,750,1000} learning_rate = {0.5,0.3,0.1,0.075,0.05,0.02,0.01} max_depth = {-1,4,5,6,7,8}
XGBoost	n_estimators = 100 learning_rate = 0.3 max_depth = 6	n_estimators = {50,100,250,500,1000,1500,2000} learning_rate = {0.5,0.3,0.2,0.1,0.075,0.05,0.02,0.01} max_depth = {4,5,6,7,8}
CatBoost	n_estimators = 1000 learning_rate = 0.03 max_depth = 6	n_estimators = {250,500,600,700,800,900,1000,1250,2000} learning_rate = {0.5,0.3,0.1,0.075,0.05,0.03,0.01} max_depth = {4,5,6,7,8}

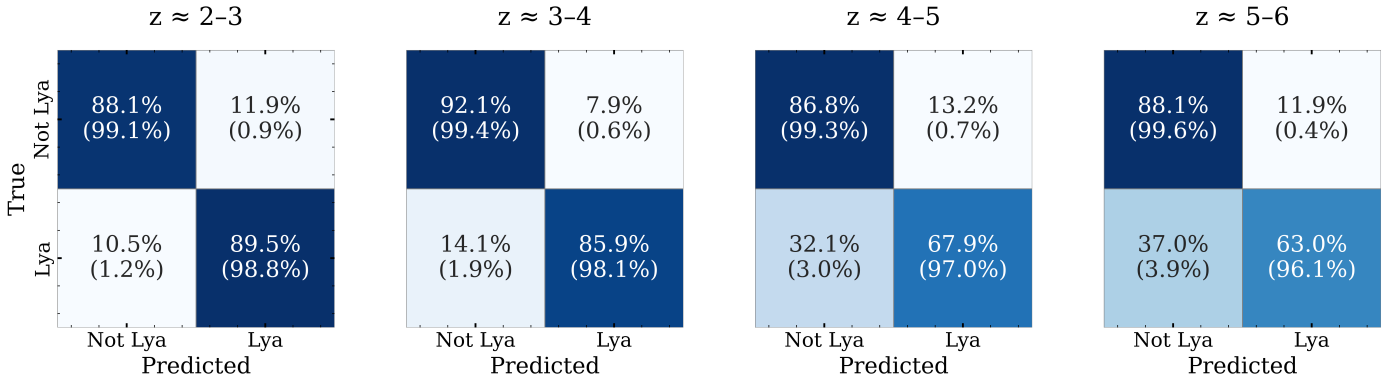
Notes. The values for max_depth of 0 or -1 result in not limited trees.

Table B.2: Final set of hyper-parameters for each algorithm and each sample, selected by the best F1-Score from GridSearch.

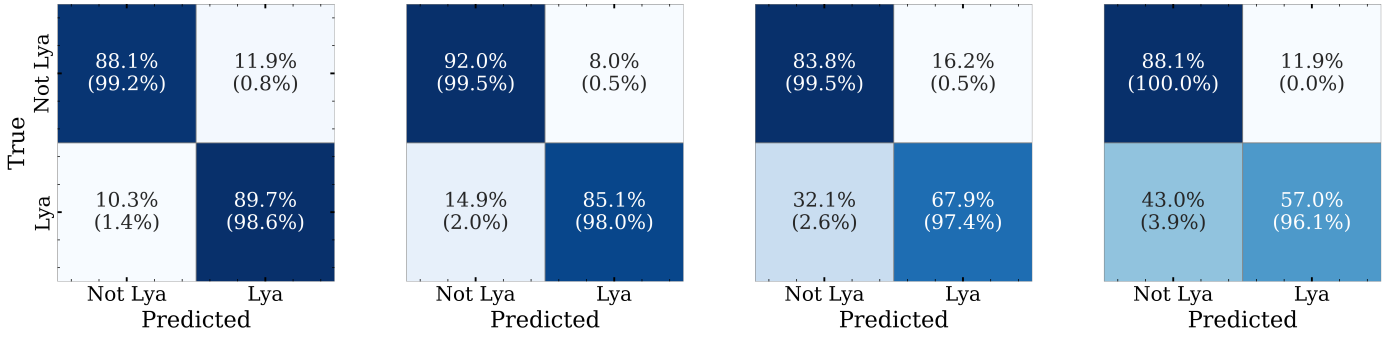
		Hyper-parameters		
Algorithm	Sample	n_estimators	learning_rate	max_depth
LightGBM	Sample 1	750	0.075	5
	Sample 2	100	0.3	7
	Sample 3	50	0.1	8
	Sample 4	500	0.05	8
	Sample 5	250	0.1	7
XGBoost	Sample 1	250	0.1	6
	Sample 2	250	0.3	8
	Sample 3	100	0.5	6
	Sample 4	1500	0.01	7
	Sample 5	250	0.075	4
CatBoost	Sample 1	1500	0.05	4
	Sample 2	2000	0.1	5
	Sample 3	1000	0.075	5
	Sample 4	100	0.05	5
	Sample 5	250	0.1	6

Appendix C: Confusion matrices separated by redshift interval

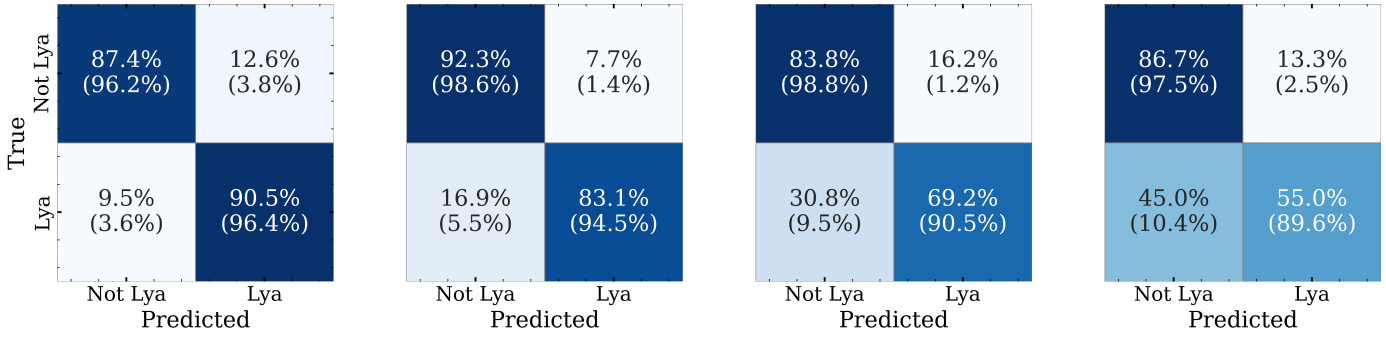
In this appendix, we present the average confusion matrices for the training and test sets across different redshift intervals: [2, 3], [3, 4], [4, 5], and [5, 6]. This is presented in Figure C.1, separately for each algorithm used. We verify that the number of true positives and true negatives is reduced as the redshift increases, leading to a poorer performance. This effect is more noticeable at $z > 4$ and is consistent across the three algorithms.



(a)



(b)



(c)

Fig. C.1: Average confusion matrices for the training (in brackets) and test sets across different redshift intervals using LightGBM (a), XGBoost (b), and CatBoost (c).

Appendix D: Spectroscopically confirmed predicted LAEs

Table D.1: Identification of the spectroscopically confirmed LAEs that were predicted by our models. The source compilation used was Khostovan et al. (2025). For each source, we present their Cosmos ID, the COSMOS2020 right ascension and declination in degrees, the spectroscopic redshift from the corresponding survey, the average prediction score attributed by our models and the corresponding reference of the observing program.

ID	RA	DEC	z	Prediction score	Reference
234035	149.93480	1.63806	4.270	0.64	Hasinger et al. (2018)
364107	150.28945	1.76708	3.536	0.93	Hasinger et al. (2018)
882240	149.96797	2.25817	4.825	0.83	Hasinger et al. (2018)
1015255	150.34350	2.38053	4.865	0.66	Hasinger et al. (2018)
1063117	149.96235	2.42651	4.331	0.53	Hasinger et al. (2018)
1065656	149.86276	2.42895	4.252	0.92	Hasinger et al. (2018)
1145690	150.32957	2.50640	4.375	0.55	Hasinger et al. (2018)
1364946	149.90115	2.71935	4.417	0.75	Hasinger et al. (2018)
935357	150.13882	2.30700	3.002	0.51	Le Fèvre et al. (2015); Tasca et al. (2017)
926044	150.10998	2.29865	2.579	0.52	Le Fèvre et al. (2015); Tasca et al. (2017)
1027774	150.11771	2.39110	3.685	0.68	Le Fèvre et al. (2015); Tasca et al. (2017)
986789	150.17796	2.35369	2.663	0.76	Le Fèvre et al. (2015); Tasca et al. (2017)
1186357	150.19970	2.54533	3.99	0.80	Le Fèvre et al. (2015); Tasca et al. (2017)
1194293	150.17779	2.55310	2.931	0.86	Le Fèvre et al. (2015); Tasca et al. (2017)
1189474	150.11301	2.54828	2.475	0.75	Le Fèvre et al. (2015); Tasca et al. (2017)
960881	150.12266	2.33115	3.100	0.83	Schmidt et al. (2021)
855981	150.11998	2.23466	5.245	0.52	Schmidt et al. (2021)
851264	150.16004	2.23055	3.362	0.50	Schmidt et al. (2021)
840426	150.10451	2.22100	5.306	0.77	Schmidt et al. (2021)
835220	150.16006	2.21626	3.009	0.69	Schmidt et al. (2021)
833340	150.12719	2.21475	3.393	0.96	Schmidt et al. (2021)
830930	150.17496	2.21229	2.973	0.53	Schmidt et al. (2021)
830219	150.09378	2.21195	3.393	0.80	Schmidt et al. (2021)
828510	150.10382	2.21038	3.244	0.94	Schmidt et al. (2021)
820571	150.12293	2.20322	4.97	0.86	Schmidt et al. (2021)
841033	150.09003	2.22157	3.247	0.81	Rosani et al. (2020)
862699	150.15868	2.24061	3.613	0.81	Rosani et al. (2020)
963145	150.11921	2.33318	3.006	0.63	Rosani et al. (2020)
297187	149.64846	1.70356	5.755	0.56	Ning et al. (2020)
897108	150.10767	2.27187	5.858	0.77	Pentericci et al. (2018)
709082	149.91040	2.10059	2.602	0.98	Lee et al. (2018); Horowitz et al. (2022)
746299	149.91977	2.13525	2.414	0.73	Lee et al. (2018); Horowitz et al. (2022)
746990	149.89142	2.13569	2.885	0.99	Lee et al. (2018); Horowitz et al. (2022)
772090	150.36777	2.15830	2.614	0.87	Lee et al. (2018); Horowitz et al. (2022)
779372	150.12211	2.16507	2.986	0.61	Lee et al. (2018); Horowitz et al. (2022)
780988	150.26258	2.16598	2.601	0.77	Lee et al. (2018); Horowitz et al. (2022)

Continued on next page

Table D.1. Continued.

ID	RA	DEC	z	Prediction score	Reference
818249	150.20273	2.20032	2.628	0.99	Lee et al. (2018); Horowitz et al. (2022)
831491	150.03077	2.21253	2.710	0.98	Lee et al. (2018); Horowitz et al. (2022)
1005589	150.22114	2.37094	2.517	0.71	Lee et al. (2018); Horowitz et al. (2022)
1014763	149.98645	2.37882	2.545	0.96	Lee et al. (2018); Horowitz et al. (2022)
1072458	149.89117	2.43519	2.682	0.70	Lee et al. (2018); Horowitz et al. (2022)
1087513	149.95876	2.45026	2.627	0.99	Lee et al. (2018); Horowitz et al. (2022)
1107108	150.03117	2.46885	2.540	0.83	Lee et al. (2018); Horowitz et al. (2022)
1115349	150.33797	2.47537	2.403	0.67	Lee et al. (2018); Horowitz et al. (2022)
1127921	150.14896	2.48828	2.380	0.60	Lee et al. (2018); Horowitz et al. (2022)
1001949	150.12686	2.36819	2.675	0.99	The HETDEX collaboration et al. (2023)
882398	150.19893	2.25778	2.611	0.99	The HETDEX collaboration et al. (2023)
876293	150.27823	2.25241	2.611	0.99	The HETDEX collaboration et al. (2023)
902799	150.24821	2.27721	2.880	0.97	The HETDEX collaboration et al. (2023)
861537	150.22581	2.23847	2.947	0.95	The HETDEX collaboration et al. (2023)
919305	150.26890	2.29190	2.763	0.92	The HETDEX collaboration et al. (2023)
920040	150.11345	2.29203	2.289	0.91	The HETDEX collaboration et al. (2023)
852094	150.11932	2.23086	2.886	0.88	The HETDEX collaboration et al. (2023)
1000124	150.26582	2.36609	2.552	0.86	The HETDEX collaboration et al. (2023)
869170	150.24103	2.24515	3.444	0.84	The HETDEX collaboration et al. (2023)
846881	150.20089	2.22575	2.481	0.84	The HETDEX collaboration et al. (2023)
904304	150.11306	2.27872	2.138	0.80	The HETDEX collaboration et al. (2023)
875173	150.09854	2.25180	3.173	0.78	The HETDEX collaboration et al. (2023)
975976	150.20227	2.34395	3.326	0.68	The HETDEX collaboration et al. (2023)
817254	150.14991	2.19838	2.698	0.50	The HETDEX collaboration et al. (2023)

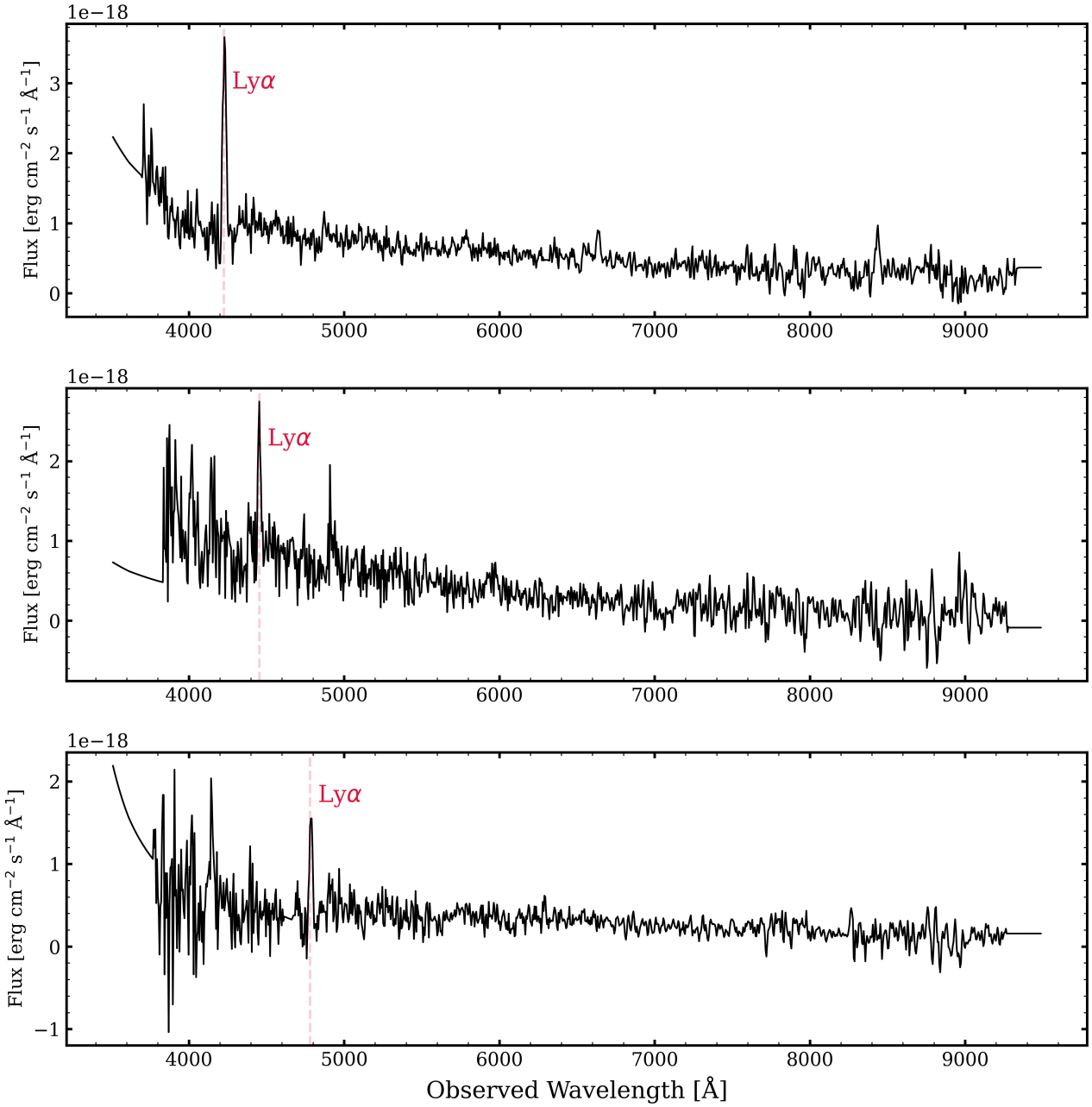


Fig. D.1: Validation spectra for three sources from Table D.1 with IDs 1189474 (top), 986789 (middle), and 1194293 (bottom).

Appendix E: Feature importance

In this appendix, we present the ten most important features, averaged across the five samples, for all three algorithms (Figure E.1). Feature importance is the process by which algorithms assign importance to each feature, allowing the user to obtain insights about the most important features for the models to make the decisions. It is assessed differently across the models. LightGBM and XGBoost use a split-based method, while CatBoost relies on how much a prediction changes when a feature value is altered.

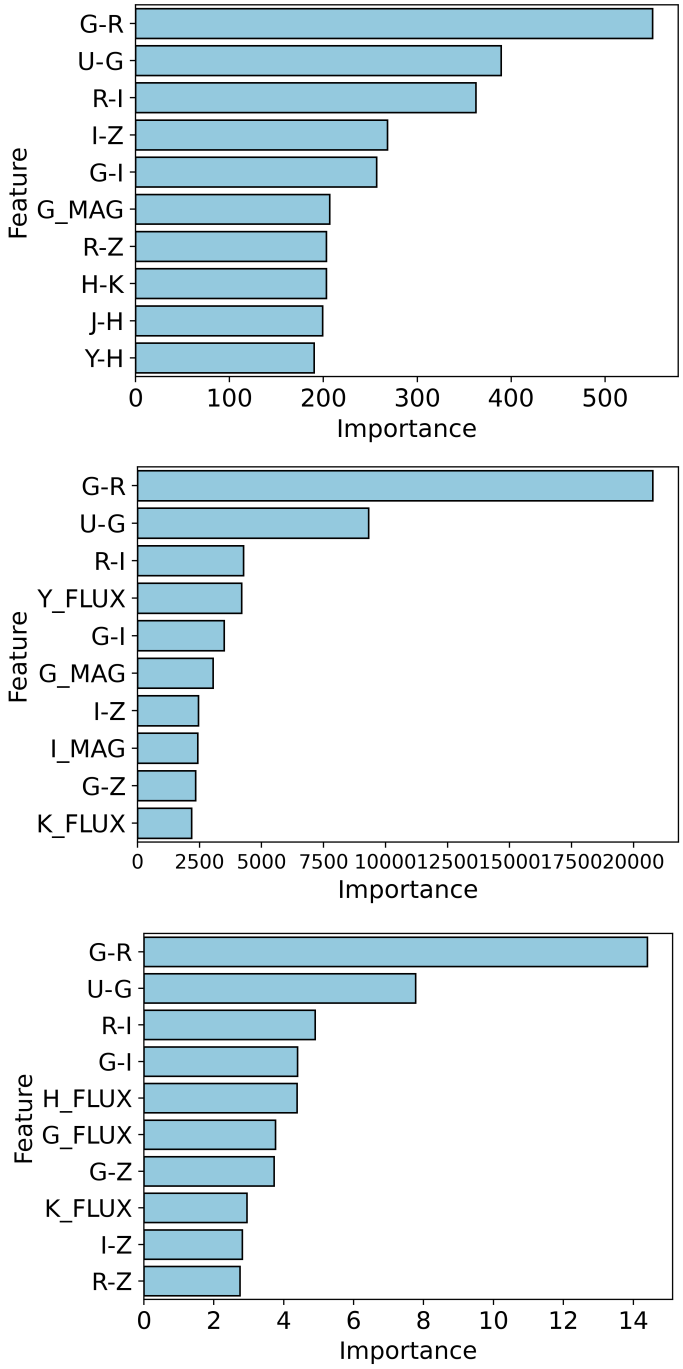


Fig. E.1: Average of the feature importances of the 10 most relevant features for LightGBM (top), XGBoost (middle), and CatBoost (bottom).