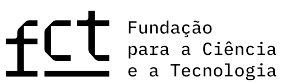
Boosting LAEs: Identification and Characterisation

Afonso Vale[1]

[1]FCUP & IA-UPorto, Portugal

April 21, 2023

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 in the form of an exploratory project with the EXPL/FIS-AST/1085/2021 reference (PI: Paulino-Afonso).



Introduction

- Can we achieve this only using broadbands in the optical and NIR?

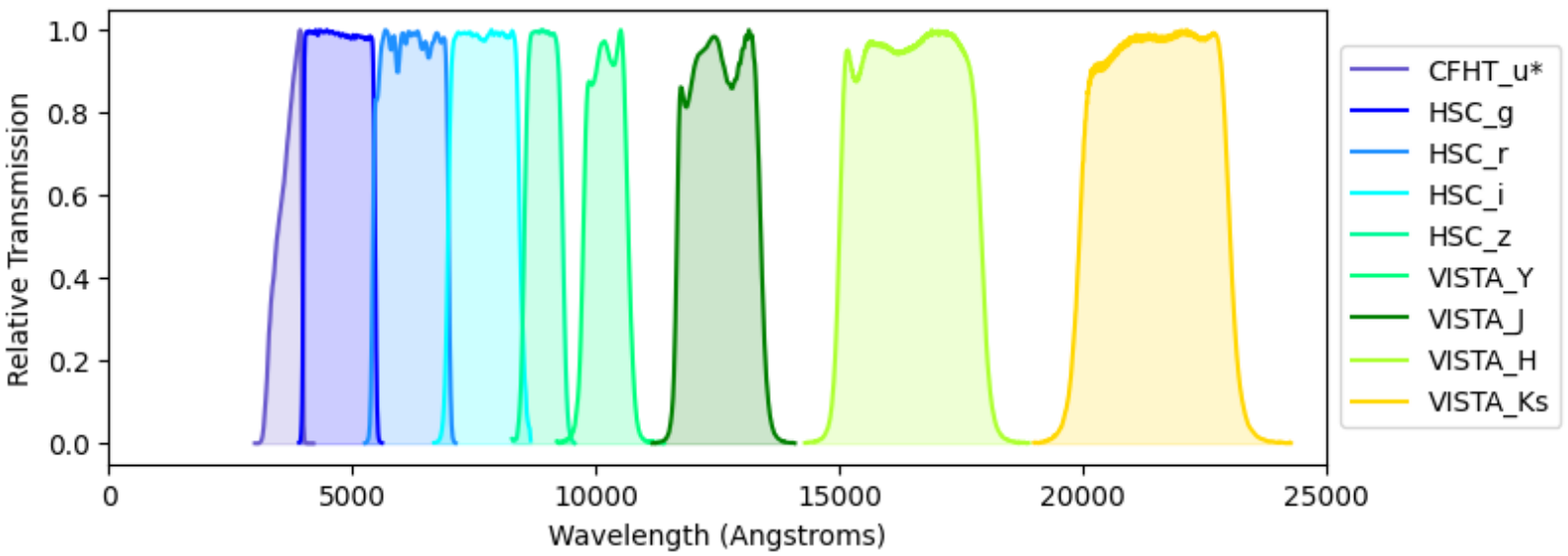


Figure 1: Relative transmission curves for the photometric bands used.

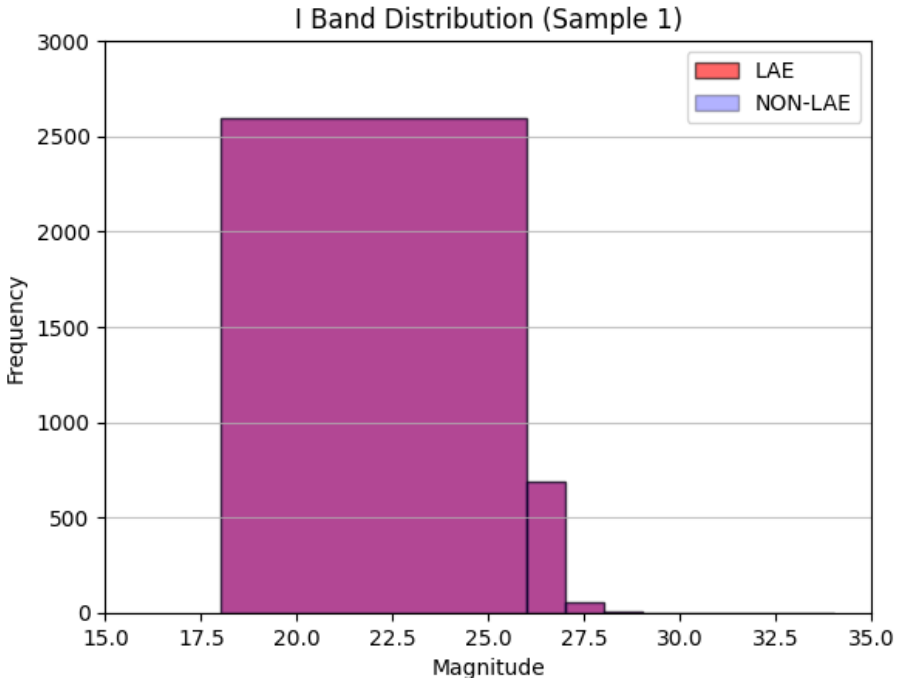
- To create this model we work with tabular data:
  -> COSMOS2020 (J. R. Weaver et al. 2021);
  -> SC4K (Sobral et al. 2018).

Data Preparation and Calibration

1. Match SC4K with COSMOS2020;
2. Restrict the i-band magnitude of the matched sample and of COSMOS2020;
3. Restrict the redshift of COSMOS2020;
4. Remove the matched sample from COSMOS2020;
5. We finally have our two samples:
   -> **Non-Lae sample** with 196199 sources (from 1.7 million);
   -> **Lae sample** with 3346 sources (originally SC4K has 3908).

Samples used in ML

- We extract 5 different subsamples of non-laes that mimick the redshift and i-band distribution of the lae sample.



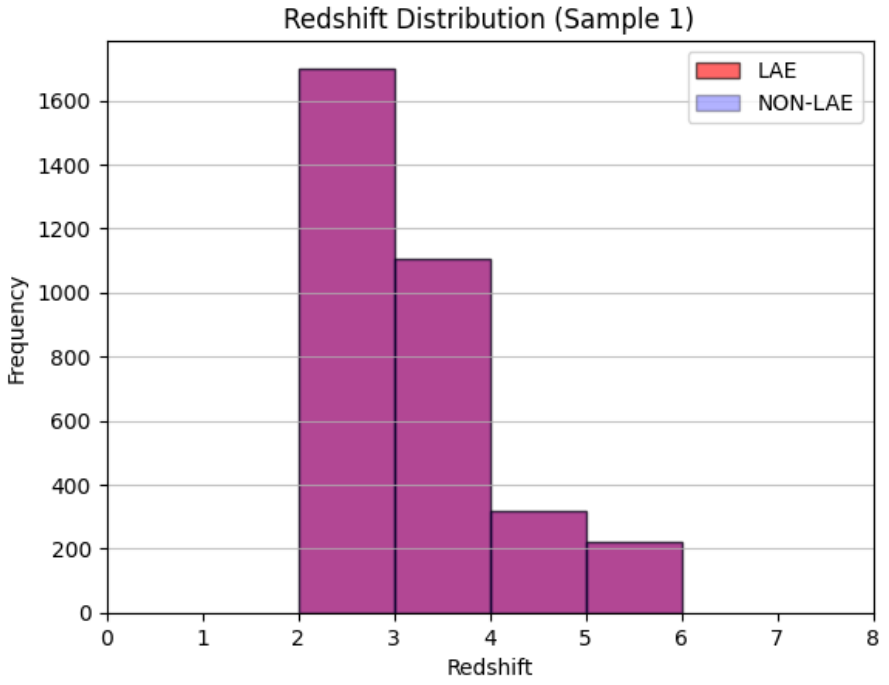


Figure 2: I-band and redshift distribution of a sample.

Comparison of Photometric Properties & ML Algortihms

Features	Algorithm	F1-Score (Train)	F1-Score (Test)	Running Time
Colors	LGBM	0.988	0.817	5s
	XgBoost	1	0.813	15s
	CatBoost	0.955	0.821	40s
Magnitudes	LGBM	0.962	0.802	5s
	XgBoost	0.999	0.802	10s
	CatBoost	0.927	0.802	20s
Fluxes	LGBM	0.958	0.826	5s
	XgBoost	0.995	0.825	10s
	CatBoost	0.930	0.833	28s
Fluxes + Colors	LGBM	0.995	0.866	20s
	XgBoost	1	0.864	18s
	CatBoost	0.974	0.868	50s
Magnitudes + Colors	LGBM	0.993	0.862	16s
	XgBoost	1	0.859	24s
	CatBoost	0.969	0.866	52s
Fluxes + Magnitudes	LGBM	0.967	0.813	5s
	XgBoost	0.998	0.813	15s
	CatBoost	0.936	0.823	26s
Mag + Colors + Fluxes	LGBM	0.994	0.866	14s
	XgBoost	1	0.857	26s
	CatBoost	0.969	0.869	49s

Figure 3: Comparison of metrics between features used.

Classification Task and Predictions

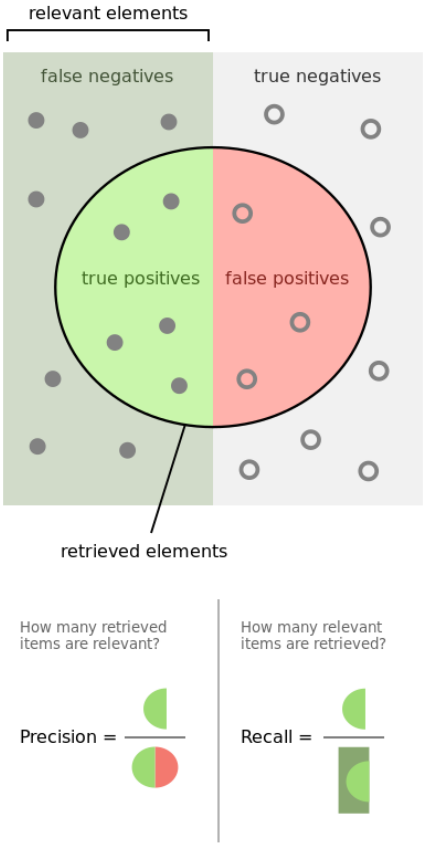


Figure 4: Walber, CC BY-SA 4.0.

	Algorithms		
	LGBM	**XgBoost**	**CatBoost**
Accuracy	0.867	0.858	0.868
Precision	0.891	0.881	0.890
Recall	0.835	0.827	0.838
F1-Score	0.866	0.864	0.868

Table 1: Summary of the evaluation metrics of Classification.

- We have 15 different models with which we can predict in the rest of the data.
- And combining their predictions:

We predict 6261 new LAEs in COSMOS2020!

Crossmatch with HEDTEX

- Crossmatching our predictions with the HEDTEX spectra catalog we get 40 matches:

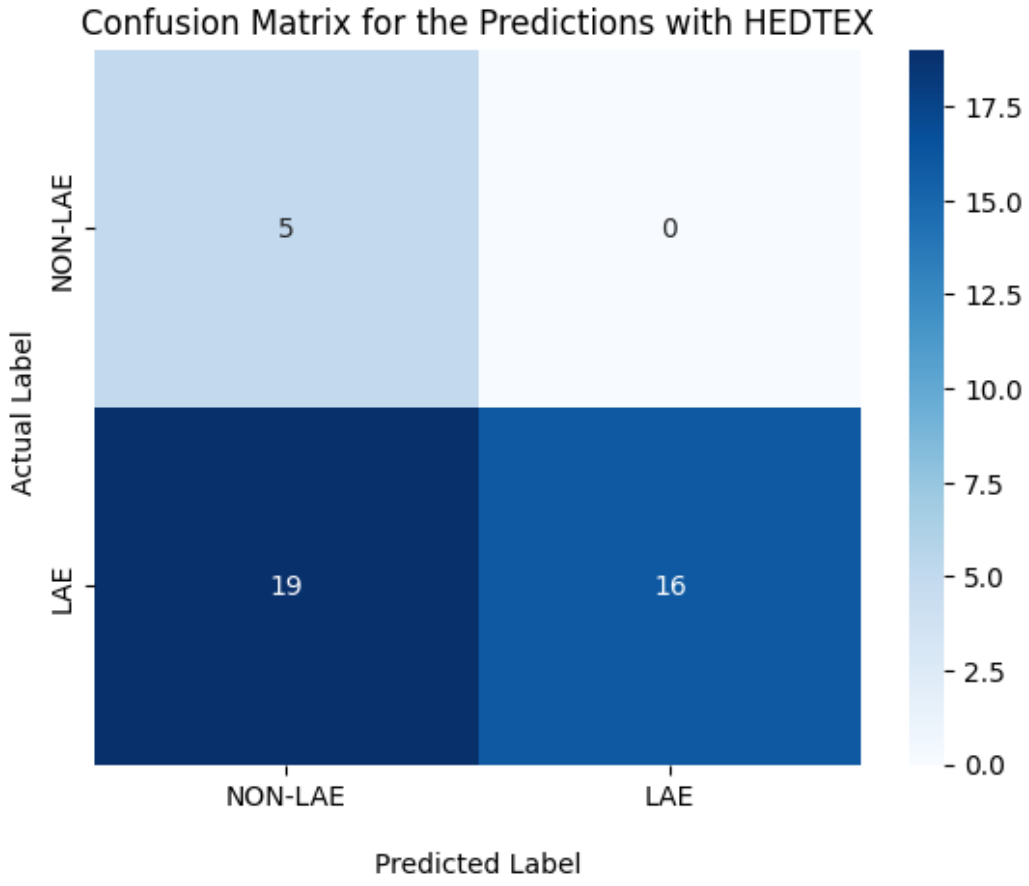


Figure 5: Confusion Matrix of our Predictions in HEDTEX.

Regression: Overview

- Only used SC4K to train and test the models;
- The same ML algorithms but suited for regression;
- Combine the predictions of each algorithm using **VotingRegressor**.

	Metrics		
	MAE	**RMSE**	**R^2**
Redshift	0.140	0.213	0.928
Lyα Luminosity	0.132	0.184	0.556
Equivalent Width	0.505	0.707	0.454

Table 2: Summary of the evaluation metrics of Regression task.

Prediction of the Redshift

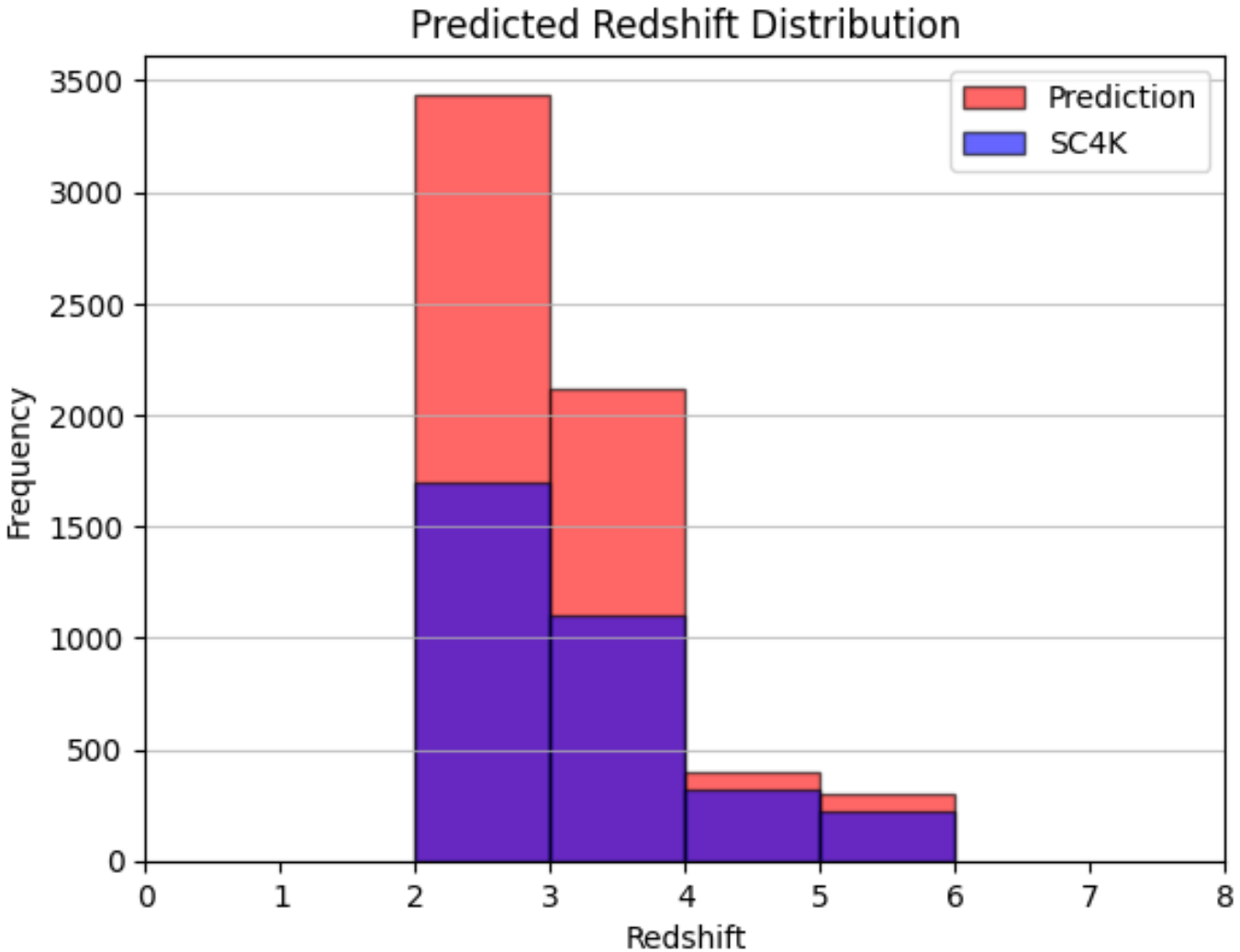


Figure 6: Prediction of the redshift.

Prediction of the Lyα Luminosity

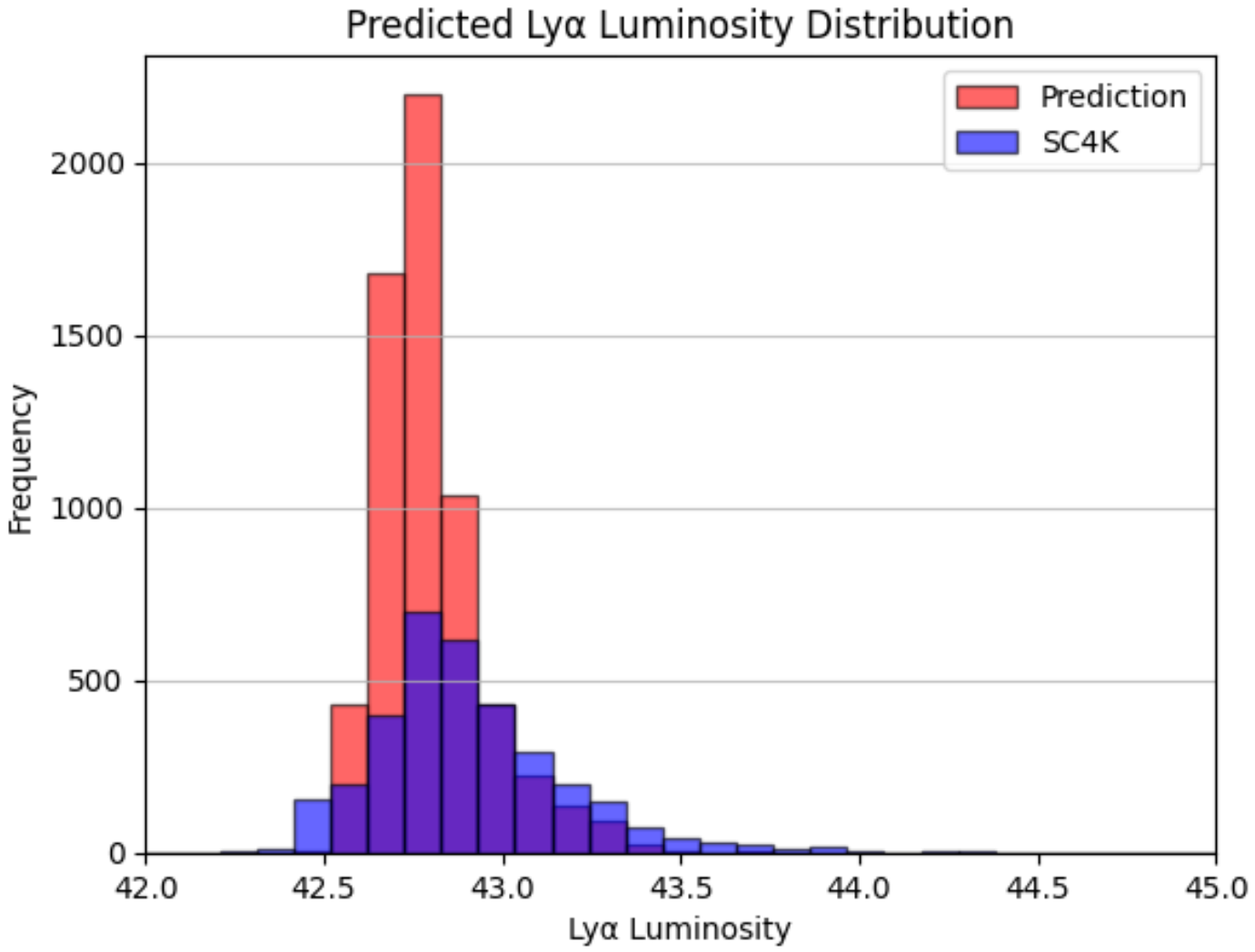


Figure 7: Prediction of the Lyα Luminosity.

Prediction of the Equivalent Width

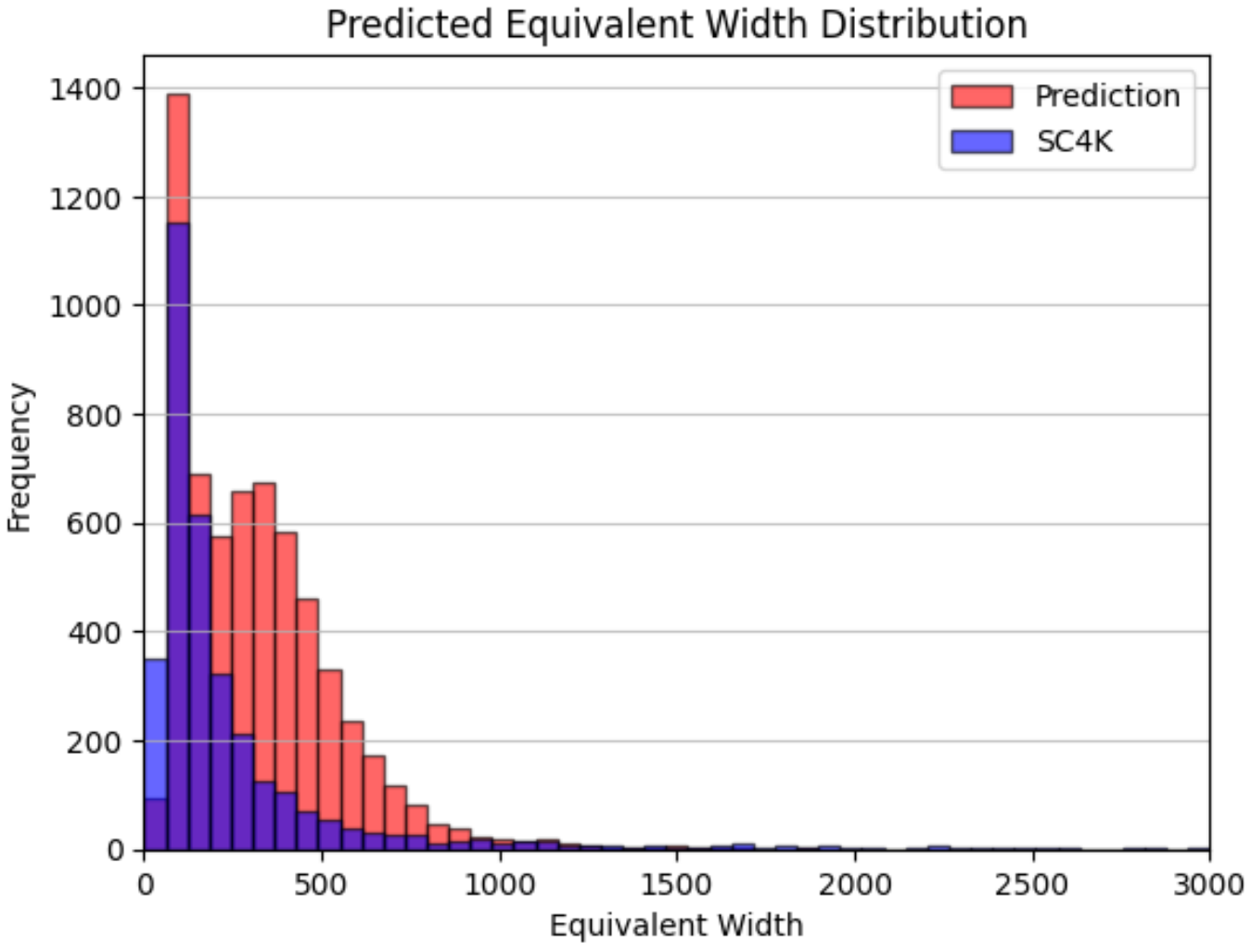


Figure 8: Prediction of the Equivalent Width.

Conclusion and Future Work

It is possible to identify and characterise LAEs using only broadband photometry!

$\hookrightarrow$ And we are able to obtain spectroscopic confirmation for some of them.

- **Future Work:**

  -> Tune the models of both tasks to achieve more polished results;

  -> Gather spectroscopic confirmation for more predicted LAEs;

  -> Generalize them to apply on other fields and larger surveys (e.g. Euclid and LSST).