# Leveraging Machine Learning for Photometric Redshift Estimation of JWST Galaxies

Julie Kalná*[1], Ryan Begley†[1], Callum Donnan[1]

[1] School of Physics and Astronomy, University of Edinburgh

**Open Access**

**Received**
20 Sep 2024

**Revised**
16 Oct 2024

**Accepted**
21 Oct 2024

**Published**
24 Oct 2024

**Abstract**

With the launch of JWST, the volume and complexity of astronomical data are increasing, a trend that will continue with future instruments such as SKA and Euclid. It is inevitable that data-driven methods will become more prominent alongside model-driven analysis. This research utilises machine learning, specifically a kernelised local linear regression model, in photometric redshift predictions of galaxies observed with JWST. With a spectroscopic dataset of 4605 galaxies, we trained the model and achieved a deviation of $\sigma_{\mathrm{dz}} = 0.018$ and a catastrophic outlier rate of $f_{\mathrm{outlier}} = 3.5\%$. These results demonstrate high accuracy and computational efficiency, highlighting the potential of machine learning for astronomical data analysis, in particular for large-scale surveys.

**DOI**: 10.2218/esjs.9996 **ISSN** 3049-7930

## Introduction

Astronomical redshift measures the shift in light wavelengths due to the expansion of the universe, providing key information about galaxy distances and recession velocities, as well as insight into the evolution and large-scale structure of the universe. There are two primary approaches taken for measuring redshift: spectroscopic redshift ($z_{\mathrm{spec}}$) and photometric redshift ($z_{\mathrm{phot}}$). Spectroscopic redshifts are obtained by taking the explicit spectrum of light from an object and comparing prominent emission lines in the spectrum to known rest-frame wavelengths. These measurements are very precise with small uncertainties (e.g., $\delta z \lessapprox 0.01$); however, they can be time-consuming and resource-intensive, especially with large-scale surveys. On the other hand, photometric techniques simply require measurements of an object's flux taken through multiple broadband filters. Some common methods used for photometric redshift estimations are spectral energy distribution (SED) fitting (Brammer *et al.* 2008) and the Lyman-break technique (Dunlop 2012). These techniques enable efficient measurements on large sets of data, often used to compile samples of galaxies for spectroscopic follow-up observations.

The James Webb Space Telescope (JWST) has provided data that looks further back in time than previously possible, allowing some of the fundamental questions about the universe to be addressed. By operating primarily in the infrared, objects from the early universe whose light has been reddened due to cosmological expansion are now detectable. JWST also offers enhanced resolution and sensitivity, enabling deeper observations of fainter, distant objects (McElwain *et al.* 2023; Wang 2024). JWST is providing vast sets of high-quality data ready to be analysed, for which machine learning can offer insights into complicated patterns that may be hard to identify otherwise (Baron 2019).

This research focuses on the implementation of a supervised machine learning model to allow scalability of $z_{\mathrm{phot}}$ measurements for galaxies observed by JWST. The goal is to develop a data-driven approach that is efficient and accurate in handling large datasets as well as free from model-dependent biases (Hainline *et al.* 2024). This research demonstrates the potential for enhancing and accelerating data analysis methods during the big data era of astronomy (Zhang *et al.* 2015).

In this work, we first provide a brief background on machine learning and its applications in redshift astronomy. We then describe the observational data and the machine learning model used in this study

*Student Author
†Corresponding academic contact: rbeg@roe.ac.uk

before presenting and discussing the results.

## Machine Learning Background

The central idea in supervised machine learning involves a mapping function, $f$, that relates feature vectors, $X$, to assigned labels, $Y$:

$$X \xrightarrow{f} Y \tag{1}$$

Because $f$ is unknown and often not rudimentary, we must derive an estimated function, $f_{\text{est}}$, such that $f_{\text{est}} \sim f$. This is done by finding a function, $f_{\text{data}}$, for a training dataset $(X_{\text{data}}, Y_{\text{data}})$ where $Y_{\text{data}}$ has known values. By then assuming that $f_{\text{est}} \sim f_{\text{data}}$, any data point that has a feature vector $\mathbf{x} \in X$ can now be assigned a label $\mathbf{y} \in Y$.

The estimated function depends on parameters, $\alpha$, that define the model. To find optimal parameters for $\alpha$, an objective function $J(\alpha)$ is defined which quantifies how well the model's predicted labels match the true labels and is typically composed of two parts:

$$J(\alpha) = L(\alpha, X_{\text{data}}, Y_{\text{data}}) + \lambda R(\alpha) \tag{2}$$

The loss term $L(\alpha, X_{\text{data}}, Y_{\text{data}})$ measures the discrepancy between the predicted labels and the true labels. The regularisation term $R(\alpha)$ penalises model complexity to prevent overfitting.

The hyperparameter $\lambda$ balances the trade-off between the terms. By minimising $J(\alpha)$ with respect to $\alpha$, we aim to find the parameters that result in the smallest error on the training data while maintaining the model's ability to generalise to new, unseen data. The model performance is assessed on a testing data set which is withheld during training.

Machine learning approaches aimed at inferring redshift estimates from photometry have previously been implemented to assist in obtaining $z_{\text{phot}}$ of galaxies in other large astronomical surveys such as the Sloan Digital Sky Survey (SDSS) (Beck *et al.* 2016). Various approaches have been utilised, including artificial neural networks (Reis *et al.* 2012; Brescia *et al.* 2014) and random forests (Carliles *et al.* 2010). These models are trained on large spectroscopic datasets to predict the redshift of objects based solely on their photometric properties. A local linear regression model was used for low redshifts in the Pan-STARRS1 survey and achieved a standard deviation of $\sigma_{\text{dz}} = 0.0299$ and an outlier rate of $f_{\text{outlier}} = 4.30\%$ (Tarrío *et al.* 2020; a definition of these metrics is provided in the 'Results' section).

However, linear regression can be limiting when trying to fit data that may not necessarily be linearly separable. This work therefore extends the approach taken in Tarrío *et al.* (2020) so that it can be applied to higher redshifts relevant to galaxies detected by JWST by adapting the local linear regression and implementing a kernel function. At higher redshifts, galaxies exhibit more complex spectral energy distributions due to factors such as evolution in galaxy properties and the effects of cosmic expansion (Dunlop 2012). This leads to non-linear relationships between photometric colours and redshift, necessitating more sophisticated models to capture these complexities. Kernel functions are used to help capture non-linear relationships in data by mapping the data to a higher dimensional feature space where non-linear data may become linearly separable (e.g., Hofmann *et al.* 2008).

## Method

In this study, a kernelised local linear regression model was used. In the context of redshift, a local linear regression model identifies galaxies close to each other in the colour space and fits the model to the nearest neighbours. The kernel function was implemented to help with capturing more complex, non-linear relationships between the JWST galaxies.

The galaxy photometry used to train the model originates from two surveys; PRIMER (e.g., Dunlop *et al.* 2021) and JADES (e.g., Eisenstein *et al.* 2023; Rieke *et al.* 2023a). From these surveys we use catalogues spanning three fields; PRIMER/UDS, PRIMER/COSMOS and JADES/GOODS-S (the reader is referred to Begley *et al.* (2024) for a detailed outline of these catalogues). The input data, $S_n$, consisted of magnitudes taken from 8 JWST NIRCam filters (F090W, F115W, F150W, F200W, F277W,

F356W, F410M, F444W; Rieke *et al.* 2023b) and 3 Hubble Space Telescope (HST) HST/ACS WFC filters (F435W, F606W, F814W; Clampin *et al.* 2000).

For all consecutive filters, a colour was calculated by taking the difference of magnitudes between the adjacent filters, defining a 10-dimensional feature space consisting of feature vectors $\mathbf{x}^{(t)} \in \mathbb{R}^{10}$ for the model to operate in. The labels are defined by the $z_{\text{spec}}$ of each galaxy, $\mathbf{y}^{(t)} \in \mathbb{R}$. Thus, the input data for the model is $S_n = \{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) \mid t = 1, \ldots, n\}$, where $n = 4605$ galaxies. The full set of $n$ galaxies is split by a 4:1 ratio into training and testing sets, ensuring stratified splitting such that the training and testing data groups are representative of the overall data.

The model operates in a local space by identifying $k$-nearest neighbours of a data point by taking Euclidean distances in the normalised 10-dimensional colour space. This newly defined decision boundary is used for fitting the model at that data point. The kernel mapping function used for the model was a Gaussian radial basis function (RBF), which is standard in kernelised regression due to its ability to handle non-linear relationships:

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2) \tag{3}$$

where $x$ and $x'$ are feature vectors and $\gamma$ is a hyperparameter that defines the width of the RBF. This implicitly computes an inner product in an infinite-dimensional feature space without explicitly computing the coordinates. Now calculating the kernel weights, $\alpha$:

$$\alpha = (K + \lambda I)^{-1} \mathbf{y} \tag{4}$$

where $\lambda$ is the regularisation hyperparameter introduced earlier, $I$ is the identity matrix, and $y$ is the vector of labels allows for the prediction. It follows that the predicted $z_{\text{phot}}$ is:

$$z_{\text{phot}} = \sum_{i=1}^{k} \alpha_i K(\mathbf{x}, \mathbf{x}_i) \tag{5}$$

where the hyperparameter $k$ is the number of nearest neighbours. The model refines the set of nearest neighbours by excluding those with large residuals based on prediction errors, $\delta_{z_{\text{phot}}}$:

$$\delta_{z_{\text{phot}}} = \sqrt{\frac{\sum_{j=1}^{k} \left( z_{\text{spec}}^{(j)} - (\alpha K)^{(j)} \right)^2}{k}} \tag{6}$$

During this refinement neighbors with residuals greater than $3 \times \delta_{z_{\text{phot}}}$ are excluded. The computation for $\alpha$ is then repeated with the remaining $l < k$ galaxies. This ensures that outliers or poorly matching data points do not negatively affect the prediction.

Hyperparameters were tuned through Bayesian optimisation (e.g., Snoek *et al.* 2012; Shahriari 2016) with a five-fold cross-validation strategy (Stone 1974) to ensure the model generalises well to unseen data. Bayesian optimisation utilises an acquisition function to determine the next point within the hyperparameter space to search by balancing exploration and exploitation. This reduces the number of evaluations required for optimisation and hence improves the model's performance without excessive computational cost. In this model, there are in total three hyperparameters that were tuned: $\lambda$, $\gamma$, and $k$.

The model was optimised on a random subset of training data and then run ten times with constant optimised hyperparameters using different random seeds to test for robustness. Each run included 921 testing galaxies, resulting from the 4:1 split of the total 4605 galaxies.

## Results

The model's performance was evaluated using three key metrics: mean squared error (MSE), standard deviation of redshift deviations, $\sigma_{\text{dz}}$, and the fraction of catastrophic outliers, $f_{\text{outlier}}$. The standard deviation $\sigma_{\text{dz}}$ was calculated based on the normalised median absolute deviation (NMAD) as $\sigma_{\text{dz}} =$

$1.4826 \times \sigma_{\mathrm{MAD}}(dz)$, where $dz = (z_{\mathrm{phot}} - z_{\mathrm{spec}})/(1 + z_{\mathrm{spec}})$. The factor 1.4826 scales the median absolute deviation (MAD) to be align with the standard deviation under the assumption of a normal distribution. The fraction of catastrophic outliers was measured as the fraction of galaxies with $|dz| > 0.15$, as typically used (Hildebrandt *et al.* 2010; Jones *et al.* 2020).

On the test set, the average $\sigma_{\mathrm{dz}}$ over the ten trials reached 0.018, with small deviations. The catastrophic outlier rate reached $f_{\mathrm{outlier}} = 3.5\%$.

In Figure 1, we present the results of our model. The top plot shows the normalised redshift deviation $dz$ as a function of the galaxy's $z_{\mathrm{spec}}$. Most points cluster around $dz = 0$, indicating good agreement between the predicted $z_{\mathrm{phot}}$ and the true $z_{\mathrm{spec}}$. The bottom plot displays the predicted $z_{\mathrm{phot}}$ versus the $z_{\mathrm{spec}}$, showing that the predictions closely follow the ideal $z_{\mathrm{phot}} = z_{\mathrm{spec}}$ line.

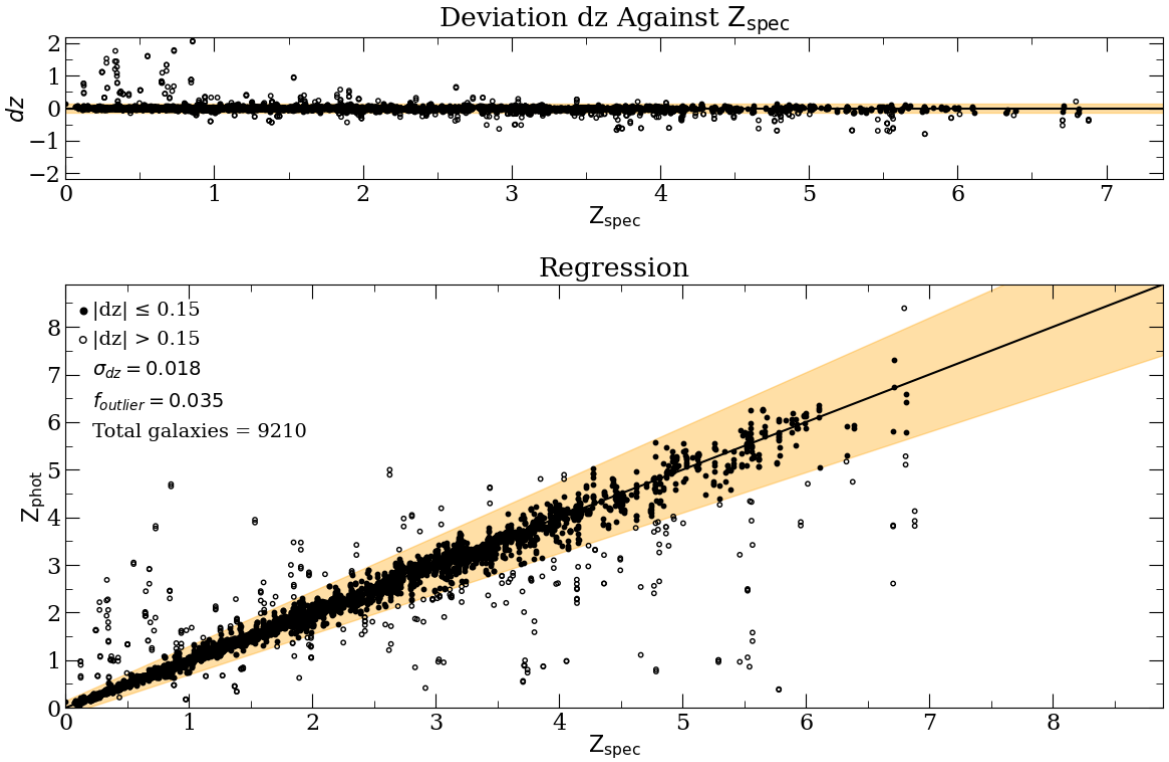


Figure 1: **Top**: The deviation of $dz = (z_{\mathrm{phot}} - z_{\mathrm{spec}})/(1 + z_{\mathrm{spec}})$ as a function of the galaxy $z_{\mathrm{spec}}$. **Bottom**: The regression plot of $z_{\mathrm{phot}}$ versus $z_{\mathrm{spec}}$. Both plots contain data from the 10 runs with varying random seeds, each run containing a total of 921 testing galaxies. The solid circles mark galaxies within the defined threshold of $|dz| < 0.15$, while the hollow circles mark galaxies outside this range. The threshold for the outliers is also shaded in orange on both plots.

These results are comparable to individual SED runs from existing methods of code for obtaining $z_{\mathrm{phot}}$ and competitive with the latest model-based $z_{\mathrm{phot}}$ measurements (Wang *et al.* 2023; Hainline *et al.* 2024).

## Discussion

The low value of $\sigma_{\mathrm{dz}} = 0.018$ shows that the model's predictions are tightly clustered around the actual values, and the overall prediction error is low. Further evaluation of the model's performance was carried out by examining the probability density function (PDF) of the normalised redshift error $dz = (z_{\mathrm{phot}} - z_{\mathrm{spec}})/(\delta_{z_{\mathrm{phot}}})$. Ideally, the model's errors would be normally distributed, closely following a standard Gaussian curve, indicating that the model's prediction errors are well-behaved and consistent with statistical expectations. There was a slight bias in the scaled data with the probability trailing off towards negative values along with a few extreme outliers which were sigma clipped. This suggests that the estimated errors $\delta_{z_{\mathrm{phot}}}$ for $z_{\mathrm{phot}}$ might need to be reevaluated on whether they give a fair assessment of the estimation accuracy, especially if the PDF were not to improve for a larger data set.

An immediate improvement in the robustness and achieved accuracies of the model could be gained from a larger spectroscopic training sample. Such improvements are imminent with ongoing spectroscopic observations from JWST. This would increase the training size which, in turn, should continue to improve the evaluation metrics as suggested by the learning curve of the model. In the meantime, there are various other steps to be taken, for example evolving the model to utilise subsampling, allowing the model to train and predict on galaxies with missing features, similar to the approach taken in Tarrío *et al.* (2020). This would increase the training data size and result in a more versatile model with fewer restraints placed on the data.

This project offers a baseline for utilising machine learning in handling large-scale datasets efficiently which will be vital for future JWST surveys such as CosmosWeb as well as for next generation large-scale surveys like Euclid or LSST.

## Conclusion

This paper demonstrated the potential of supervised machine learning in predicting $z_{\rm phot}$ with both accuracy and computational efficiency. Using a kernelised local linear regression model we achieve a standard deviation, $\sigma_{\rm dz} = 0.018$, and catastrophic outlier rate, $f_{\rm outlier} = 3.5\%$, on a relatively sparse spectroscopic data set of 4605 galaxies. The input vectors for the model were constructed in a 10-dimensional colour space using flux data from JWST and HST surveys. Hyperparameters were tuned using Bayesian optimisation and a 5-fold cross validation strategy was implemented to ensure generalisability of the model. An RBF kernel was used to allow for linear regression of non-linearly separable data. The paper also explored potential further modifications of the method which may improve its capabilities such as refining the model to handle missing data. It should be expected for the model to only improve from here as more data becomes available. This research demonstrates the potential of further applications with similar machine learning models in large astronomical surveys, not just limited to JWST, but possibly also for various upcoming large-scale surveys.

## Acknowledgments

The author wishes to thank and acknowledge the support of Dr. Ryan Begley and Callum Donnan who supervised the project. The project was funded by the Carnegie Trust for the Universities of Scotland.

## Software Availability

The code written for this project is available at:
`https://github.com/Juli-Kalna/JWST-ML-Galaxy-Photo-Z`

## References

Baron, D. 'Machine Learning in Astronomy: A Practical Overview' arXiv e-prints (2019)

Beck, R. *et al.* 'Photometric Redshifts for the SDSS Data Release 12' Monthly Notices of the Royal Astronomical Society **460** 2 (2016)

Begley, R. *et al.* 'The Evolution of [OIII]+H$\beta$ Equivalent Width From z$\simeq$ 3 − 8: Implications for the Production and Escape of Ionizing Photons During Reionization' arXiv e-prints (2024)

Brammer, G. B. *et al.* 'EAZY: A Fast, Public Photometric Redshift Code' The Astrophysical Journal **686** 2 (2008)

Brescia, M. *et al.* 'A Catalogue of Photometric Redshifts for the SDSS-DR9 Galaxies' Astronomy & Astrophysics **568** (2014)

Carliles, S. *et al.* 'Random Forests for Photometric Redshifts' The Astrophysical Journal **712** 1 (2010)

Clampin, M. *et al.* 'The Advanced Camera for Surveys' in UV, Optical, and IR Space Telescopes and Instruments (International Society for Optics and Photonics; 2000)

Dunlop, J. S. *et al.* 'PRIMER: Public Release IMaging for Extragalactic Research' *JWST Proposal Cycle 1* (2021)

Dunlop, J. S. 'Observing the First Galaxies' in The First Galaxies: Theoretical Predictions and Observational Clues (Springer Berlin Heidelberg; 2012)

Eisenstein, D. J. *et al.* 'Overview of the JWST Advanced Deep Extragalactic Survey (JADES)' arXiv e-prints (2023)
Hainline, K. N. *et al.* 'The Cosmos in Its Infancy: JADES Galaxy Candidates at Z $>$ 8 in GOODS-S and GOODS-N' The Astrophysical Journal **964** 1 (2024)
Hildebrandt, H. *et al.* 'PHAT: PHoto-Z Accuracy Testing' Astronomy & Astrophysics **523** (2010)
Hofmann, T. *et al.* 'Kernel Methods in Machine Learning' The Annals of Statistics **36** 3 (2008)
Jones, E. and Singal, J. 'Tests of Catastrophic Outlier Prediction in Empirical Photometric Redshift Estimation With Redshift Probability Distributions' Publications of the Astronomical Society of the Pacific **132** 1008 (2020)
McElwain, M. W. *et al.* 'The James Webb Space Telescope Mission: Optical Telescope Element Design, Development, and Performance' Publications of the Astronomical Society of the Pacific **135** 1047 (2023)
Reis, R. R. R. *et al.* 'The Sloan Digital Sky Survey Co-add: A Galaxy Photometric Redshift Catalog' The Astrophysical Journal **747** 1 (2012)
Rieke, M. J. *et al.* 'JADES Initial Data Release for the Hubble Ultra Deep Field: Revealing the Faint Infrared Sky With Deep JWST NIRCam Imaging' arXiv e-prints (2023)
Rieke, M. J. *et al.* 'Performance of NIRCam on JWST in Flight' Publications of the Astronomical Society of the Pacific **135** 1044 (2023)
Shahriari, B. 'Practical Bayesian Optimization With Application to Tuning Machine Learning Algorithms' (2016)
Snoek, J. *et al.* 'Practical Bayesian Optimization of Machine Learning Algorithms' arXiv e-prints (2012)
Stone, M. 'Cross-Validatory Choice and Assessment of Statistical Predictions' Journal of the Royal Statistical Society: Series B (Methodological) **36** 2 (1974)
Tarrío, P. and Zarattini, S. 'Photometric Redshifts for the Pan-STARRS1 Survey' Astronomy & Astrophysics **642** (2020)
Wang, B. *et al.* 'UNCOVER: Illuminating the Early Universe–JWST/NIRSpec Confirmation of Z $>$ 12 Galaxies' The Astrophysical Journal Letters **957** 2 (2023)
Wang, J. 'Eye Beyond the Sky' (Springer Nature Singapore; 2024)
Zhang, Y. and Zhao, Y. 'Astronomy in the Big Data Era' Data Science Journal **14** (2015)