Machine Learning Techniques for High-Redshift Galaxy Classification with JWST NIRCam Data

Rosa Roberts, Christopher J. Conselice
Jodrell Bank Centre for Astrophysics, University of Manchester
Oxford Road, Manchester, M13 9PL, UK
rosa.roberts@student.manchester.ac.uk, conselice@manchester.ac.uk

Abstract

We present machine learning techniques to identify high-redshift galaxies using multi-band imaging from the James Webb Space Telescope (JWST). Our primary method utilises CNNs trained on galaxy cutouts taken in various JWST NIRCam filters to detect the Lyman break, achieving up to 98% accuracy in distinguishing high- from low-redshift sources. However, brown dwarfs are frequently misclassified due to their similar red spectral energy distributions, highlighting challenges in feature discrimination. We also develop a semi-supervised pipeline combining HDBSCAN clustering with a Random Forest classifier to exploit the structure of unlabelled data. HDBSCAN effectively isolates clusters corresponding to populations at $z \geq 8$ and lower redshifts, albeit with some overlap. Training a Random Forest on these clusters enables identification of $z \geq 10$ galaxies, while lower redshifts remain less distinctly separated. These scalable methods facilitate efficient identification of high-redshift galaxy candidates in large JWST surveys, complementary to traditional photometric and spectroscopic techniques.

1 Introduction

Identifying high-redshift galaxies is fundamental to advancing our understanding of cosmic structure formation and galaxy evolution in the early Universe. These distant systems provide crucial insights into key epochs such as reionisation, the formation of the first stars and galaxies, and the emergence of large-scale structure (Papovich et al., 2025; Robertson et al., 2010). The James Webb Space Telescope (JWST; McElwain et al., 2023), with its unparalleled infrared sensitivity and spatial resolution, has opened an unprecedented window into this early cosmic era, enabling the detection of galaxies at redshifts as high as $z \sim 14$ (Ferrara, 2024). However, despite the depth and richness of JWST's imaging data, reliably distinguishing high-redshift galaxies from the more numerous low-redshift interlopers remains a significant challenge (Signor et al., 2024; Hovis-Afflerbach et al., 2021; Bisigello et al., 2016).

Traditional redshift estimation techniques fall into two main categories: spectroscopic and photometric. Spectroscopic methods determine redshift by detecting emission or absorption lines in a galaxy's spectrum or by directly observing the *Lyman break*, offering highly precise redshift measurements (Frontera-Pons et al., 2019; Zhou et al., 2021). However, the significant telescope time and resources required make these observations unfeasible for the vast datasets produced by modern surveys (Jamal et al., 2018; Navarro-Gironés et al., 2024). For instance, Euclid is set to observe around 1.5 billion galaxies over six years, returning roughly 100 GB of data daily (Merlin et al., 2023). Photometric redshift estimation, based on broadband imaging, offers greater scalability but is vulnerable to systematic uncertainties, especially for faint or distant galaxies (Dahlen et al., 2013). A common technique is template-fitting, which compares observed fluxes to spectral energy distribution (SED) models (e.g. Brammer et al., 2008; Larson et al., 2023). Although widely used, this approach is computationally demanding and prone to mismatches, especially at high redshift where SED templates are less reliable. It also generally ignores morphological information, which can further limit its effectiveness (Jamal et al., 2018; Clausen et al., 2025).

The rapid growth of multi-wavelength survey data has enabled the rise of machine learning (ML) techniques for photometric redshift estimation (e.g. Zhou et al., 2021; Tanigawa et al., 2024; Pasquet et al.,

2018). These ML models learn mappings between photometric inputs and redshift labels from training datasets and offer substantial improvements in speed and scalability over traditional methods. Recent work shows that ML models can match or even surpass the accuracy of template-fitting, and they are increasingly being adopted in large-scale cosmological surveys (Henghes et al., 2021; Signor et al., 2024).

Despite their promise, ML methods have yet to be extensively validated for identifying the earliest galaxies. The scarcity of spectroscopically confirmed high-redshift sources and the limited availability of deep near-infrared imaging constrain the construction of representative training sets (Razim et al., 2021; Tanigawa et al., 2024). These limitations underscore the importance of developing ML techniques capable of harnessing the full depth of available imaging data.

Deep learning, particularly convolutional neural networks (CNNs; O'Shea & Nash, 2015), has become a powerful tool for extracting complex patterns from pixel-level data. CNNs show strong potential for identifying high-redshift galaxies, as they can extract both spatial patterns and spectral information from multi-band imaging data (e.g. Zhong et al., 2024; Cheng et al., 2020; Fu et al., 2024). Of particular interest are strong spectral discontinuities like the Lyman and Balmer breaks, which shift into the near- and mid-infrared at higher redshifts (Yue et al., 2013; Kuruvanthodi et al., 2024), offering robust signals for ML classification.

Lyman Break Galaxies (LBGs) are star-forming systems identified by a sharp flux drop at wavelengths shorter than 912Å due to neutral hydrogen absorption in the star forming regions of the galaxies (Giavalisco, 2002; Taran, 2024). LGBs also exhibit Ly-α emission at 1216Å from intervening neutral hydrogen in the interstellar medium near the central region of a galaxy (Ouchi et al., 2020). These features shift into the observable optical and near-infrared bands at high redshifts due to cosmological expansion, enabling their detection via the *dropout* technique, an approach that, while effective, remains prone to contamination (Yue et al., 2013; Kuruvanthodi et al., 2024). In multi-band imaging, the dropout effect produces sharp flux contrasts between adjacent filters, patterns that CNNs can efficiently learn (Taran, 2024). By preprocessing inputs to emphasise these features, we train a CNN to recognise the photometric signatures of early-Universe galaxies across a wide redshift range. Example SEDs illustrating the shift of the Lyman break with redshift are shown in Appendix A.

The Balmer break is a discontinuity a galaxy's SED at 3645Å arising from photon absorption by hydrogen atoms transitioning from the $n = 2$ state (Labbe et al., 2024). While the Lyman and Balmer breaks occur at distinct rest-frame wavelengths, their redshifted signatures can overlap in observed bands, leading to degeneracies in photometric redshift estimation. This often results in aliasing between low- and high-redshift solutions, manifesting as bimodal redshift probability distributions (Adams et al., 2025). To address this, we compute the expected redshift aliases from the known rest-frame positions of these breaks, enabling a more systematic interpretation of such cases.

Another source of contamination in high-redshift galaxy searches comes from brown dwarfs: substellar objects too low in mass to sustain hydrogen fusion. These cool, faint objects emit primarily in the infrared, and their SEDs can closely resemble the colours of high redshift galaxies. Specifically, they exhibit strong molecular absorption features that can mimic the Lyman break, and as unresolved point sources, they can visually resemble small, distant galaxies (Tu et al., 2025). This similarity leads to frequent misclassification, complicating identification. Distinguishing brown dwarfs from high-redshift galaxies requires careful analysis of subtle spectral and morphological differences. However, the intrinsic faintness and rarity of brown dwarfs result in limited data samples, making CNN training difficult. We therefore explore data augmentation techniques to improve classification performance.

In addition to supervised CNNs, both unsupervised learning methods and hybrid approaches that combine unsupervised and supervised techniques have demonstrated strong performance in categorising galaxy morphologies (e.g. Tohill et al., 2024; 2023; Kolesnikov et al., 2023). Recent studies indicate that these frameworks can match, and in some cases rival, traditional machine learning methods, while closely approaching the performance of deep learning models, particularly in the classification of astronomical objects (Asadi et al., 2025). However, it remains untested whether a hybrid pipeline can be applied to galaxy data for the identification of high redshift galaxies. In this study, we apply the t-SNE (van der Maaten & Hinton, 2008) and HDBSCAN (McInnes & Healy, 2017) algorithms to identify structure within the data and generate cluster-based labels, which are then used to train a Random Forest classifier (Breiman, 2001) for redshift classification, in order to assess whether a hybrid pipeline can effectively distinguish high-redshift galaxy populations. In Section 2,

we summarise the data and preprocessing steps. Section 3 details the machine learning models and methods used for galaxy classification, including CNNs and semi-supervised clustering. Results and their implications are presented in Section 4, followed by a discussion in Section 5. Finally, we summarise our conclusions in Section 6.

2 Data Sources and Sample Selection

The survey image, sourced from Adams et al. 2024, is part of the imaging data from the JWST Advanced Deep Extragalactic Survey (JADES; Eisenstein et al., 2023) Data Release 3 (DR3; D'Eugenio et al., 2024) in the GOODS-South field (Great Observatories Origins Deep Survey-South; Dickinson et al., 2002). Our focus is on data obtained in seven NIRCam (Near-Infrared Camera) wide-band filters: F090W, F115W, F150W, F200W, F277W, F356W, and F444W (Rieke et al., 2023). These filters cover a wavelength range $\sim 0.9 - 4.4\,\mu$m. We use the galaxy catalogue from Austin et al. In Prep., which contains 85,420 candidates including brown dwarf contaminants. We generate our galaxy redshift labels by running `EaZy-py` (Brammer et al., 2008) to perform traditional SED fitting using the `fsps_larson` templates (Larson et al., 2023). `EaZy-py` models observed photometry as linear combinations of synthetic galaxy SEDs to estimate a redshift probability density function (PDF). To ensure high-quality image samples, we apply selection criteria for both low- and high-redshift galaxy populations. These criteria are detailed in Austin In Prep. and briefly summarised below.

For low-redshift galaxies, we require sources to be unmasked in at least two bands from the Hubble Space Telescope's Advanced Camera for Surveys (ACS), and in at least six NIRCam bands, specifically including F090W, F277W, F356W, F410M, and F444W. A signal-to-noise ratio (SNR) greater than 8 is required in the first two wide-bands redward of the Lyman-α break (excluding F070W and F850LP), and SNR > 3 in all widebands redward of Lyman-α. We also require a reduced chi-squared value $\chi^2_\nu < 3.0$ from `EaZy-py` fitting using `fsps_larson` templates in 0.32as apertures. The photometric redshift probability distribution must have more than 60% of the total PDF within $|\Delta z|/z < 0.1$ of the peak. Only sources with best-fit photometric redshifts below the low-redshift threshold are included (Austin, In Prep.).

For high-redshift galaxies, we require sources to be unmasked in at least two ACS bands and at least six NIRCam bands, including F090W, F277W, F356W, F410M, and F444W. All bands entirely blueward of the Lyman limit ($\lambda_{\rm rest} < 912$Å) must be undetected at the 2σ level, and bands blueward of Lyman-α must have SNRs less than 3. The first two widebands redward of Lyman-α must have SNR > 8, and all other redward widebands must have SNR > 3. We require $\chi^2_\nu < 3.0$ for the best-fit high-redshift solution, which must also be significantly better than the low-redshift fit with $\Delta\chi^2 > 4.0$. The photometric redshift PDF must be well-constrained, with more than 60% of the probability density within $|\Delta z|/z < 0.1$ of the peak. Finally, we require sources to be resolved in all selection bands, with `SExtractor` (Bertin & Arnouts, 1996) half-light radii $R_{50} > 45$mas (1.5 pixels) (Austin, In Prep.).

To identify potential brown dwarf contaminants from the selected dataset, we fit a range of atmospheric models, including `Sonora Bobcat` (Marley et al., 2021), `Sonora Cholla` (Karalidi et al., 2021), `Sonora Diamondback` (Morley et al., 2024), `Sonora Elf Owl` (Mukherjee et al., 2024), `LOW-Z` (Meisner et al., 2021), and `ATMO2020` (Phillips et al., 2020). Sources with a best-fit $\chi^2_\nu < 3.0$ for any brown dwarf template are excluded from the main sample and retained for separate analysis (Harvey & Austin, 2025).

After applying our selection criteria, we retain 722 high-redshift galaxies. To construct a balanced training set, we randomly select an equal number of low-redshift galaxies from the filtered sample using `np.random.choice`. Balancing the dataset is crucial, as CNNs often struggle to generalise and yield less interpretable results for under-represented classes (Dablain et al., 2022). We assign numerical labels to three classes as follows: 0 for low-redshift galaxies, 1 for high-redshift galaxies, and 2 for brown dwarf candidates. Galaxies are classified based on their best-fit redshifts ($z_{\rm best}$) derived from `EaZy-py` fitting (Brammer et al., 2008), where $z_{\rm best} < 4$ indicates low-redshift and $z_{\rm best} \geq 4$ indicates high-redshift.

2.1 Creation of Multiband Cutouts

Using the RA and Dec coordinates from the galaxy catalogue in Austin et al. In Prep., we extract 1,444 galaxy cutouts and 11 brown dwarf cutouts from the survey images. Each cutout is sized 64×64 pixels. For each of the seven wide-band filters, we generate corresponding cutouts of the same sources, resulting in multi-channel cutouts. The set of cutouts is split into two groups, 85% in the training set and 15% in the testing set. During training, 15% of the data is set aside as a validation set to monitor performance after each epoch. Since this data is withheld from the

model during training, it serves as a proxy for the test set and helps determine when to apply early stopping. The test set, unseen throughout training, is used only after training to evaluate final model performance (Fu et al., 2024).

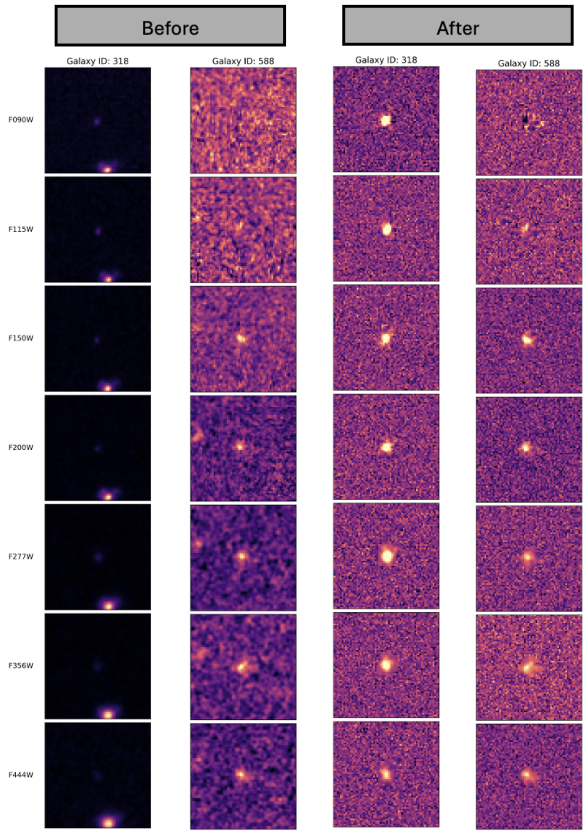


Figure 1: NIRCam cutouts of two example galaxies: one low-redshift (ID 318) and one high-redshift (ID 588), shown before and after segmentation and flux normalisation. Galaxy 588 exhibits a characteristic dropout in the F090W and F115W filters, consistent with the presence of a Lyman break.

2.2 Object Isolation via Segmentation Maps

To isolate each object from nearby contaminants, we apply segmentation maps. A segmentation map is a pixel-level mask that identifies the spatial extent of sources within an image (Xu & Zhu, 2024). Each pixel is assigned an integer label: background pixels are marked as 0, target sources with their unique ID from the catalogue, and other sources with distinct non-zero values. This allows us to retain only the target object, replacing all other labelled sources with noise generated from the background pixel distribution.

We use the segmentation map produced in Conselice et al. 2024 to apply consistent masking across all cutouts in each of the seven wide-band filters. Overlapping sources are suppressed using Gaussian noise, and any `NaN` values are similarly replaced to ensure clean input for model training. The resulting cutouts are stacked into a tensor of shape [N, 7, H, W], where N = 1,455 is the number of sources, 7 is the number of filters, and H & W denote the spatial dimensions.

2.3 Augmentation and Feature Extraction

To prevent overfitting during model training, we incorporate data augmentation into our pre-processing pipeline and apply feature extraction techniques to enhance model performance. Augmentation is particularly important when using raw pixel data, as CNNs benefit from exposure to a more diverse dataset. Without augmentation, rare or outlier examples may be under-represented and ignored during training. By artificially increasing the size and variety of the dataset, we improve the generalisability of the model and reduce the risk of overfitting (Fu et al., 2024).

Each segmented image is augmented through rotations in 10° increments from 0° to 350° anticlockwise (Cheng et al., 2020). Images are also flipped along the x- and y-axes. Gaussian noise is applied to the original, rotated, and flipped cutouts. In addition to augmentation, we apply the Histogram of Oriented Gradients (HOG; Dalal & Triggs, 2005) for feature extraction. HOG captures the distribution and directionality of local image gradients, providing structural information that complements raw pixel intensities. It has been shown to improve CNN performance in galaxy classification tasks (Cheng et al., 2020), and we evaluate its utility in enhancing model performance by applying it to both the rotated and flipped images.

This process yields several distinct training datasets: (1) the original segmented cutouts (1,055 tensors), (2) the original cutouts with Gaussian noise applied (1,055 tensors), (3) the augmented set of rotated and flipped cutouts with Gaussian noise, totalling 40,093 tensors, and (4) the HOG-transformed version of set (3), containing 40,093 tensors.

2.4 Image Normalisation and Flux Scaling

The pixel intensities of galaxy cutouts vary significantly due to differences in intrinsic brightness, morphology, and redshift. These variations can introduce bias into machine learning models, making it difficult to learn class boundaries that are invariant to brightness alone. To address this, we rescale the

pixel values of each cutout to lie within a standard range of $[0, 1]$. While intrinsic brightness can be a useful classification criteria, our primary interest lies in identifying high-redshift galaxies. In such cases, the Lyman break, rather than flux, is a more reliable indicator of redshift. Our normalisation strategy is designed to enhance contrast near this break (Cheng et al., 2020). Figure 1 shows example galaxy cutouts before and after segmentation and normalisation.

We normalise each augmented cutout based on its flux in the reddest band (F444W). We compute the minimum pixel value across the entire reddest-band cutout and the maximum value within a small central region (a 10×10-pixel box) to capture the galaxy's core signal. The full stack of 7-channel cutouts is then scaled according to these reddest-band statistics, using the following transformation:

$$\tilde{x} = \frac{x - \min(\mathrm{R})}{\max(\mathrm{R}_{\mathrm{core}}) - \min(\mathrm{R})} \tag{1}$$

where x is the pixel value in any band, $\min(\mathrm{R})$ is the minimum value in the full reddest-band cutout, and $\max(\mathrm{R}_{\mathrm{core}})$ is the maximum within the central region of that band.

3 Machine Learning Models for Galaxy Classification

3.1 Convolutional Neural Networks

CNNs (O'Shea & Nash, 2015) are a class of deep learning models particularly effective for processing image data. CNNs automatically learn spatial hierarchies of features through the use of local connections and parameter sharing.

3.1.1 Feature Extraction with Convolutional Layers

Convolutional layers are the core of CNNs, designed to extract spatial features from input datacubes. They apply learnable kernels that slide across the input to detect local patterns, generating feature maps that preserve spatial arrangement and highlight where specific patterns occur (Cheng et al., 2020). Each feature map is produced by convolving the input with a kernel, adding a bias, and applying a non-linear activation function. The first convolutional layer typically processes a multispectral input datacube, while subsequent layers take the previous layer's feature maps as input, refining and building upon earlier representations. Non-linear activation functions, particularly the Rectified Linear Unit (`ReLU`; Nair & Hinton, 2010), defined as $f(x) = \max(x, 0)$, are essential for introducing complexity and enabling the network to learn non-trivial mappings (Pasquet et al., 2018).

3.1.2 Dimensionality Reduction via Pooling Layers

Pooling layers are used to down-sample the spatial dimensions of the feature maps, typically by applying a fixed-size window (2×2 in our case) that moves across the input. This reduces the number of parameters and computations in the network, making it more efficient. By summarising regions of the feature maps, pooling also helps the model become more robust to small distortions in the input—an effect known as shift invariance. The most common form is `MAXPOOLING`, which outputs the maximum value within each window, highlighting the most prominent features. Another variant is `AVERAGEPOOLING`, which computes the mean value in the window. Pooling is typically applied after convolution and activation steps (Goodfellow et al., 2009; Pasquet et al., 2018).

3.1.3 High-Level Feature Integration with Fully Connected Layers

Fully connected (FC) layers appear near the end of the CNN and are used to transform the features extracted by the convolutional and pooling layers into a high-level representation for classification or regression tasks. In these layers, each neuron is connected to every neuron in the previous layer, enabling the network to combine and interpret features globally (Cheng et al., 2020).

3.1.4 Classification Output Layer

We formulate the problem of distinguishing high- versus low-redshift galaxies as a classification task rather than a (non-linear) regression problem. This approach has been shown to offer distinct advantages in previous studies (Wang et al., 2024; Pasquet et al., 2018). Each input multispectral image is centred on a single object, which belongs exclusively to one of three classes. To enforce this mutual exclusivity, we apply a `softmax` activation function (Bridle, 1990) to the output layer, ensuring that the class probabilities sum to one.

3.1.5 Residual Networks

Residual Networks (ResNet; He et al., 2015a) are a class of CNNs designed to address the degradation problem in very deep models. As neural networks

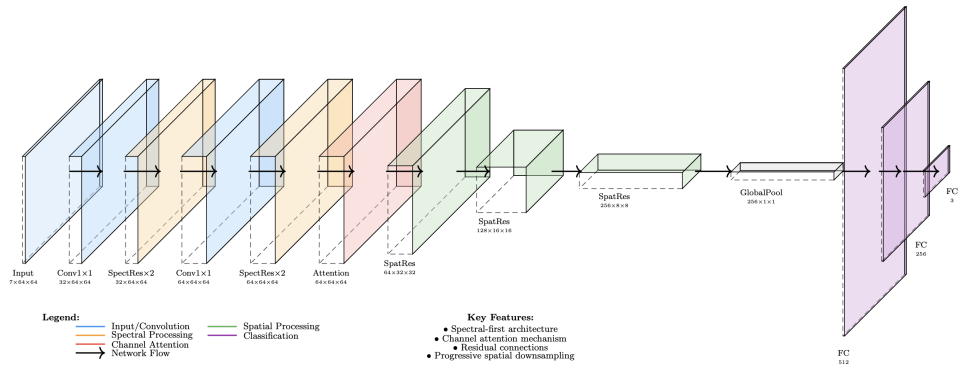


Figure 2: Architecture of the `LymanBreakClassifier`. The model begins with 1×1 convolutions and `SpectralResidualBlocks` to capture inter-band relationships, followed by `ChannelAttention` to emphasise informative spectral channels. Spatial features are then extracted via `SpatialResidualBlocks` interleaved with `MAXPOOLING` to reduce resolution while increasing depth. A global `AVERAGEPOOLING` layer aggregates features before classification is performed through a fully connected network with `BATCHNORMALISATION` and dropout.

grow in depth, they can suffer from vanishing gradients, leading to saturated or degraded training accuracy (Roy et al., 2025). ResNet mitigates this by introducing `residual (or skip) connections`, which allow the network to learn residual functions, i.e., the difference between input and output—rather than full transformations.

A standard residual block consists of two or more convolutional layers, with a shortcut connection that bypasses these layers and adds the original input x to the block's output $\mathcal{F}(x)$, resulting in $\mathcal{F}(x) + x$. This improves gradient flow and training stability, allowing deeper and more effective architectures. The modular design of ResNet enables easy integration of additional mechanisms such as attention modules or astrophysics-specific processing blocks. Its strong performance on multiband data has been demonstrated in prior studies (e.g. Bialopetravičius et al., 2019; Burke et al., 2019), making it a robust choice for our goal of identifying Lyman break signatures in large astronomical datasets.

3.1.6 Custom ResNet Architecture

We implement two custom types of residual blocks to address the distinct characteristics of spectral and spatial information in multi-spectral imaging: the `SpectralResidualBlock` and the `SpatialResidualBlock`. The `SpectralResidualBlock` consists of two 1×1 convolutional layers, each followed by `BATCHNORMALISATION` and `ReLU` activation, along with a dropout layer (Srivastava et al., 2014) for regularisation. These 1×1 convolutions operate across the spectral (channel) dimension, allowing the block to learn inter-band dependencies while preserving spatial resolution. The `BATCHNORMALISATION` layer works to normalise the input distribution during training to prevent large internal shifts (Ioffe & Szegedy, 2015; Fu et al., 2024).

The `SpatialResidualBlock` extracts spatial features using two 3×3 convolutions with `BATCHNORMALISATION`, `ReLU` activation, and spatial dropout. These kernels capture local spatial patterns such as textures and object boundaries. When downsampling or changing the number of channels, a 1×1 convolution in the skip connection ensures dimensional alignment.

3.1.7 Incorporating Channel Attention Mechanisms

To improve the model's sensitivity to the most informative spectral bands, we incorporate a channel attention mechanism based on the Squeeze-and-Excitation (SE) framework (Hu et al., 2019). This module adaptively recalibrates channel-wise feature responses by learning interdependencies between spectral channels, allowing the network to focus on the bands that contribute most to the task, such as those containing significant features like the Lyman break.

The `ChannelAttention` module operates on the output of a previous layer and emphasizes important spectral bands while suppressing less relevant ones. It begins by applying global `AVERAGEPOOLING` and

MAXPOOLING across the spatial dimensions, producing two channel descriptors. These descriptors are passed through a shared two-layer fully connected network with a reduction ratio $r = 8$, enabling the module to capture non-linear interactions between channels. The outputs are combined and passed through a sigmoid activation (Dubey et al., 2022) to generate attention weights, which are applied multiplicatively to the input tensor, enhancing or suppressing each channel accordingly.

3.1.8 Implementation Details of Our CNN Model

The CNN model developed for this analysis is implemented using the PyTorch framework (Paszke et al., 2019), and its architecture is shown in Figure 2. The network takes as input a 7-band image of size 64×64 pixels and begins with a 1×1 convolution to project the spectral bands into a higher-dimensional space. This is followed by a series of SpectralResidualBlocks that perform deep spectral feature extraction while preserving spatial structure. Channel dimensionality is expanded mid-way through this spectral processing stage to increase representational capacity. A ChannelAttention module is then applied to recalibrate the spectral features. Subsequently, the network processes the attended feature maps using stacked SpatialResidualBlocks. These extract higher-level spatial representations using 3×3 convolutions, interleaved with MAXPOOLING layers to progressively reduce spatial resolution. The number of feature channels increases across blocks, allowing the model to capture increasingly abstract spatial structures. Diagrams of the core architectural components are provided in Appendix B.

After spatial processing, a global AVERAGEPOOLING layer compresses the feature maps to a fixed-length vector, which is passed through a fully connected classifier. This classifier consists of two dense layers with BATCHNORMALISATION, ReLU activations, and dropout regularisation, followed by a final output layer with three logits for classification. Dropout rates are tuned to reduce overfitting, especially important given the relatively small size of labelled astronomical datasets.

Weight initialisation follows the He normal (He et al., 2015b) scheme for convolutional layers, with BATCHNORMALISATION and linear layer biases set to zero, ensuring stable and efficient training. The network was trained using the cross-entropy loss function (Solla et al., 1988), which is widely used for multi-class classification tasks as it measures the negative log-likelihood of the true class. Optimisation was performed using Adam, the Adaptive Moment Estimation algorithm (Kingma & Ba, 2017), chosen for its ability to adapt learning rates during training and its strong performance on deep, multi-branch architectures. A learning rate of 0.0005 was selected to strike a balance between training stability and convergence speed.

3.2 Semi-Supervised Learning Pipeline

To evaluate the reliability of clustering methods for classifying our data, we implement a semi-supervised pipeline inspired by Asadi et al. 2025, with the implementation details outlined below.

3.2.1 Visualising High-Dimensional Data with t-SNE

T-Distributed Stochastic Neighbour Embedding (t-SNE) is a powerful non-linear dimensionality reduction technique primarily used to visualise high-dimensional data in two or three dimensions. Its main goal is to preserve the local structure of the data by placing similar data points close together in the low-dimensional embedding space, thereby revealing meaningful clusters and patterns that may be difficult to detect otherwise (van der Maaten & Hinton, 2008). In our analysis, we construct a feature set comprising galaxy magnitudes and flux radii across all bands. After normalising the data using scikit-learn's StandardScaler, we apply the TSNE algorithm from scikit-learn (Pedregosa et al., 2018) to project the high-dimensional feature space into two dimensions, enabling visualisation of the intrinsic structure of the dataset.

3.2.2 Clustering with HDBSCAN

The HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm (McInnes & Healy, 2017) is a density-based clustering method that excels at detecting clusters of varying densities and shapes, while also effectively identifying noise points. Using a minimum cluster size of 5, we apply the HDBSCAN algorithm from scikit-learn to the scaled feature data, to assign cluster labels for each galaxy. To visualise the clustering results, labels are plotted over the two-dimensional TSNE embedding of the data. Figure 3 shows the TSNE projection coloured by $z_{\rm best}$, the F090W-band magnitude and the HDBSCAN cluster labels. This plot highlights distinct groups in the feature space, with some data points labelled as noise (indicated by the label -1). The clustering separates galaxies with the highest redshifts ($\gtrsim 10$) from the rest of the dataset.

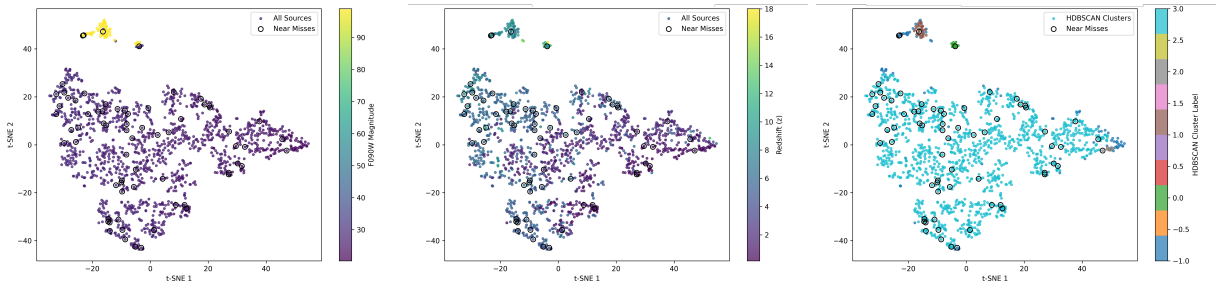


Figure 3: t-SNE projections coloured by: (a) photometric redshift (z_{best}), (b) F090W-band magnitude, and (c) HDBSCAN cluster labels. Bimodal galaxies are circled. HDBSCAN clusters galaxies with $z_{\text{best}} \gtrsim 8$, which t-SNE separates from the rest of the data. Points points labelled as -1 are identified as 'noise'.

3.2.3 Random Forest Classifier

Random Forest (RF) is a widely used ensemble learning method that aggregates the predictions of multiple decision trees to produce a final classification (Breiman, 2001; Cheng et al., 2020). Each tree in the forest is trained on a randomly selected subset of the training data using bootstrap sampling, and each split within a tree considers a random subset of the input features. This randomness encourages diversity among the trees, which improves generalisation and reduces overfitting (Fawagreh et al., 2014).

In this work, we use the `RandomForestClassifier` from the `scikit-learn` library. The model consists of 100 decision trees, each grown without a predefined maximum depth, allowing them to expand until all leaves are pure or no further splits are possible. For each node split, the number of features considered is set to the square root of the total number of input features. The final classification decision is made by majority voting across all trees. Random forests have been shown to be effective in a range of astronomical classification tasks, particularly when working with parameter inputs (derived numerical features rather than raw pixel data) (Dubath et al., 2011; Beck et al., 2018).

4 Results

4.1 Evaluation factors for the models

To evaluate model performance we use the Receiver Operating Characteristic (ROC) curve (Fawcett, 2006; Powers, 2020), implemented via `scikit-learn`. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR), where an ideal classifier curves toward the top-left corner (0, 1), indicating high sensitivity and low false alarm rate. The TPR and FPR are defined as:

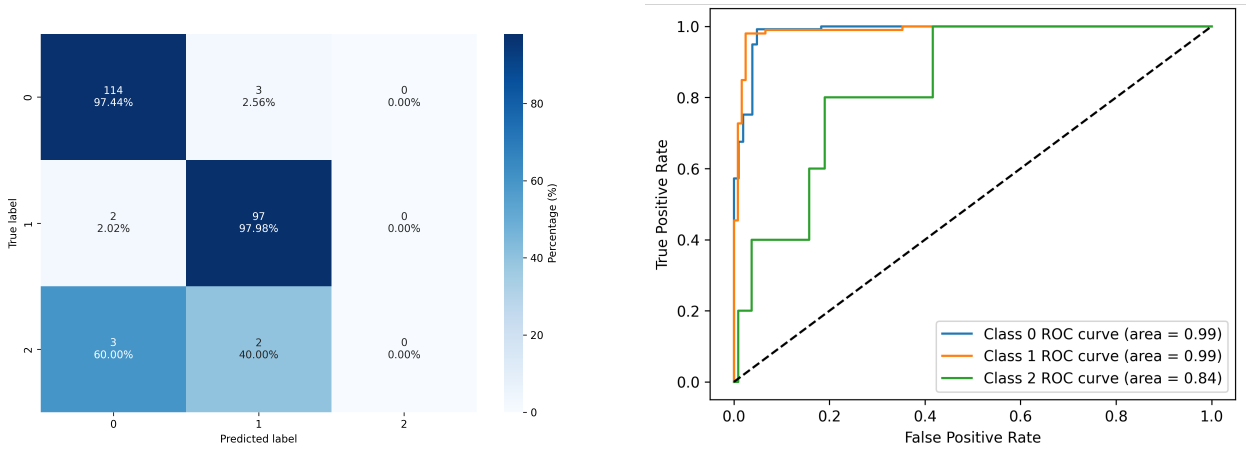
$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} \quad = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{2}$$

We also compute the area under the ROC curve (`scikit-learn`'s `AUC` method) as a scalar performance metric (Fawcett, 2006). The `AUC` represents the probability that the classifier ranks a randomly chosen positive sample higher than a randomly chosen negative one. It serves as an indicator how well the model distinguishes between the classes (Cheng et al., 2020).

Confusion matrices are used to visualise the proportions of TP, FP, TN, and FN, offering an interpretable summary of the model's classification performance. To monitor training behaviour and implement early stopping, we plot both training and validation loss across epochs to identify overfitting and assess convergence. Additionally, we track training and validation accuracy, defined as the proportion of correctly classified samples relative to the total number of predictions:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{3}$$

This provides a direct measure of overall model correctness during training and evaluation (Cheng et al., 2020). Clustering performance for algorithms such as `HDBSCAN` and `TSNE` is evaluated using `scikit-learn`'s `Silhouette Score` method, which quantifies how well each data point fits within its assigned cluster compared to other clusters. It is calculated using the mean intra-cluster distance and the mean nearest-cluster distance, with scores ranging from -1 (poor clustering) to 1 (well-separated clusters). Scores near 0 indicate overlapping clusters (Shahapure et al., 2020).



(a) Confusion matrix.

(b) ROC curve.

Figure 4: Performance of the best-performing CNN run. (a) The model correctly classified 97.44% of low-redshift (0) and 97.98% of high-redshift (1) galaxies, but misclassified all brown dwarfs (2) as galaxies. (b) The ROC curve shows AUC scores of 0.99 for both galaxy classes and 0.84 for brown dwarfs.

4.2 CNN Classification Performance

After 10 training epochs ($\sim$ 10 minutes on a GPU), the CNN model reliably distinguishes between high- and low-redshift galaxies. When trained on the original dataset supplemented with augmentations (2) and (3), and evaluated on non-augmented test data, it achieved classification accuracies of $\sim 97\%$ for low-redshift and $\sim 98\%$ for high-redshift galaxies. These results were consistent across 10 independent training runs. Figures 4a and 4b illustrate the model's performance on a representative test run.

The model consistently failed to identify brown dwarfs. All brown dwarfs were misclassified as either low- or high-redshift galaxies. In the majority of runs, $\sim 60\%$ were labelled as low-redshift. The multi-class ROC curves reflect this: the model achieved AUC scores of $\sim$ 0.97 for both galaxy classes, but only $\sim 0.6 - 0.8$ for brown dwarfs. The accuracy consistently reached $\sim 95-97\%$ by the end of each training epoch, meaning the model rapidly converged and was able to generalise well across the galaxy classes. However, this overall accuracy is somewhat misleading, as it is dominated by the high performance on the more numerous low- and high-redshift galaxies. Including the HOG-augmented dataset (4) in the training set produced the same results. This indicates that the model's performance is already saturated with the features provided by the original and earlier augmented datasets, and that HOG-based augmentation does not provide additional discriminative power for this classification task.

4.3 Photometric Band Importance via Ablation Study

To quantify the relative importance of each photometric band in the CNN's classification performance, we perform an ablation analysis in which each band is systematically removed from the input, and the resulting drop in classification accuracy is recorded. This procedure is repeated across multiple independent training runs to ensure robustness of the ranking.

Across all experiments, the F200W band consistently emerges as the most critical for accurate classification, with average performance drops exceeding 45% upon its removal. This may be partially explained by the fact that F200W is often the least noisy filter, potentially making it more reliable for the model. F150W, F277W, and F356W also show substantial importance, with typical drops ranging from 30–45%, depending on the run. The F444W band shows intermediate importance, with performance drops in the 30–40% range across most runs, suggesting it provides complementary information but is less essential than the core bands. Interestingly, F090W and F115W show lower importance despite often covering the Lyman break for many galaxies—an effect that may reflect their higher noise levels.

Overall, the band importance ranking stabilises across runs as follows (in descending order of importance): F200W, F277W, F150W, F356W, F444W, F115W, and F090W. This ordering reflects the

model's reliance on mid-infrared bands, which are most effective in distinguishing high-redshift galaxies from low-redshift galaxies. These bands span wavelengths that bracket the redshifted Lyman and Balmer breaks, providing critical spectral leverage. For future model development and observational strategies, prioritising filter coverage in the 2–4µm range appears essential for robust redshift classification.

4.4 Semi-supervised Redshift Estimation

The t-SNE projection reveals a separation of very high-redshift galaxies ($z_{\rm best} \geq 8$) in feature space, particularly those with strong dropouts in the F090W band. Colouring by $z_{\rm best}$ confirms that this region corresponds to the highest-redshift galaxies in the dataset, although there is some overlap with lower-redshift sources. HDBSCAN identifies these same high-redshift systems as a distinct cluster and further isolates a smaller subgroup with $z_{\rm best} \geq 10$, including a few sources with $z_{\rm best}$ exceeding 14. While these $z_{\rm best}$ values are likely overestimated, their grouping suggests consistent photometric patterns linked to extreme redshifts. HDBSCAN also forms a separate cluster of low-redshift ($z_{\rm best} < 2$) galaxies with low F090W magnitudes, and labels some sources as noise, though this occasionally includes real objects. Silhouette scores of ~ 0.5 reflect moderate separation between clusters, implying some overlap in feature space.

To test how well the HDBSCAN clustering aligns with redshift structure, we train a RF classifier on the HDBSCAN labels and use it to predict the classes of galaxies in the test set. Binning the data into four redshift intervals (<4, 4–8, 8–10, and >10) for evaluation, the resulting confusion matrix (Figure 5) shows that the model reliably isolates galaxies with $z_{\rm best} \gtrsim 10$. However, objects in the 8–10 range are distributed across multiple clusters, mirroring the ambiguity in the t-SNE structure. Galaxies in the 4–8 bin are largely grouped with those at lower redshifts, suggesting their photometric features lack sufficient distinction for separation with these inputs alone.

4.5 Analysis of Bimodal Photometric Redshift Distributions

To evaluate the performance of both supervised and semi-supervised pipelines, we select a subset of galaxies from a clean sample of 1,444 brown dwarf-free galaxies that exhibit bimodal redshift PDFs. These bimodal PDFs, derived from `EaZy-py` fitting (Brammer et al., 2008) using `fsps_larson` templates (Larson et al., 2023), feature a secondary peak indicative of potential confusion between low- and high-redshift solutions. Example PDFs from this subset are shown in Appendix C.

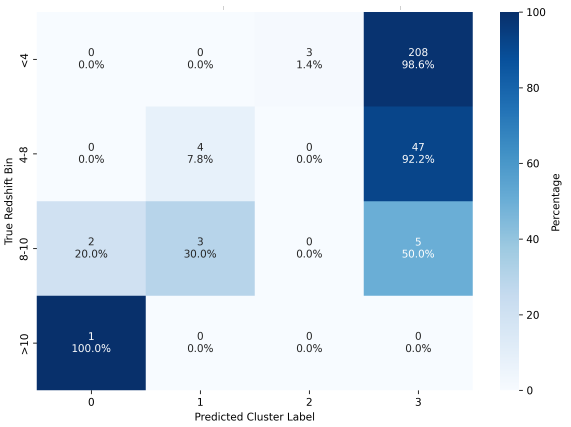


Figure 5: Confusion matrix from a representative run of the RF classifier trained on HDBSCAN-derived cluster labels. The matrix illustrates the model's classification performance across all redshift bins, highlighting areas of agreement and misclassification between predicted and true cluster assignments.

For each galaxy, we normalise its PDF and identify all peaks with heights at least 50% of the primary peak's height, where the primary peak corresponds to $\texttt{z}_{\texttt{best}}$. To assess whether a significant secondary peak corresponds to a plausible alternate redshift solution, we use the Balmer break relation to estimate the expected redshift of the secondary peak.

For high-redshift galaxies ($z_{\rm best} > 4$), we compute the expected low-redshift alias as:

$$z_{\rm low} = \left(\frac{\lambda_{\rm low}}{\lambda_{\rm high}}\right)(1 + z_{\rm high}) - 1. \tag{4}$$

Conversely, for low-redshift sources, we solve for the high-redshift alias:

$$z_{\rm high} = \left(\frac{\lambda_{\rm high}}{\lambda_{\rm low}}\right)(1 + z_{\rm low}) - 1, \tag{5}$$

where $\lambda_{\rm low} = 0.1216$ μm (Dominique et al., 2018) and $\lambda_{\rm high} = 0.3645$ μm (Wilkins et al., 2023) correspond to the Balmer break bounds.

A galaxy is flagged as bimodal if it contains a secondary peak located within $\Delta z = \pm 0.5$ of the expected alias redshift. We identify 67 galaxies whose

PDFs meet the bimodal criteria. These cases typically reflect ambiguous redshift solutions that could challenge classification accuracy. We remove these 67 galaxies from the training and validation sets, then retrain the CNN without them. The retrained model is subsequently evaluated on cutouts of the previously excluded galaxies. The resulting confusion matrix is shown in Figure 6. The CNN demonstrates strong performance, achieving a TPR of 99.9% for low-redshift galaxies and a TNR of 96.2% for high-redshift galaxies. These results are consistent with the model's overall performance on the broader test set, suggesting strong generalisation even on ambiguous cases.

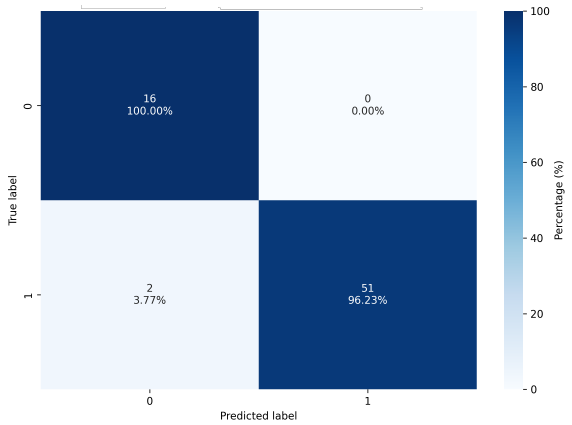


Figure 6: Confusion matrix for the CNN evaluated on cutouts of the bimodal galaxies. The model achieves a TPR of 99.99% for low-z_{best} galaxies and a TNR of 96.23% for high-z_{best} galaxies.

To investigate their behaviour within the semi-supervised pipeline, we examine the positions of these galaxies in the `t`-SNE and `HDBSCAN` representations; they are shown as circles in Figure 3. Both models fail to isolate the bimodal galaxies from the broader dataset: the 67 sources appear scattered throughout the visualisations, with no evidence of distinct clustering. This indicates that photometric structure alone, specifically magnitude and size measurements, is insufficient to resolve redshift degeneracies in bimodal systems, especially in the absence of spectral information. Consequently, unsupervised clustering approaches used to select high-confidence training labels overlook such complex edge cases. Future improvements would include incorporating additional features such as morphological indicators to better distinguish these degenerate sources.

5 Discussion

5.1 Limitations of the CNN Architecture

There are several limitations of the architecture that should be noted. The design separates spectral and spatial processing stages, extracting spectral features initially and only later integrating spatial information. This approach limits the network's ability to jointly model spectral-spatial interactions at early stages, potentially missing subtle patterns where spatial and spectral features are tightly coupled.

By default, the CNN assigns each input to the class with the highest predicted probability. Introducing a confidence threshold ($p > 0.8$) results in a slight reduction in classification completeness, as some predictions are now withheld due to lower confidence. For low-redshift galaxies, $\sim 97\%$ are correctly classified, around 3% are deemed ambiguous. High-redshift galaxies show slightly lower confidence: $\sim 85-88\%$ are correctly classified, $\sim 10-14\%$ fall into the ambiguous category, and $\sim 1-2\%$ are misclassified as low-redshift. Brown dwarfs are predominantly misclassified as galaxies, with $\sim 60\%$ being assigned to the low-redshift galaxy class. However, in some instances, one brown dwarf is classified as ambiguous when applying the confidence threshold. Results from two representative threshold runs are presented in Appendix D.

While the `ChannelAttention` module effectively recalibrates spectral channels, it does not incorporate spatial attention. Spatial attention mechanisms can further improve performance by focusing the network's resources on spatial regions containing relevant astrophysical sources, a valuable addition given the localised nature of Lyman Break galaxies in images. Additionally, the spectral blocks do not explicitly encode positional or physical relationships between spectral bands. Given that spectral bands correspond to ordered wavelengths with known astrophysical significance, the absence of positional encoding or mechanisms to exploit this domain knowledge may constrain the model's ability to fully leverage spectral structure.

This model was trained without formal hyperparameter optimisation. Parameters such as dropout rates, number of channels, learning rates, and layer depths were chosen heuristically rather than through systematic tuning. Consequently, the architecture's performance can be further improved by targeted hyperparameter search and architectural refinements.

5.2 Model Limitations in Brown Dwarf Classification

Our analysis shows that the limited number of brown dwarf samples significantly hinders the CNN's ability to distinguish them from high-redshift galaxy cutouts. We begin with only 11 brown dwarf cutout tensors, which we augment to 429 and further expand to 858 using HOG transformations. However, brown dwarfs are typically unresolved point sources with limited spatial structure, rendering augmentations such as rotations largely ineffective in introducing meaningful variability, since a rotated point source remains visually unchanged. Moreover, this dataset size remains negligible compared to the much larger non-brown dwarf training set.

Despite augmentation, the model fails to learn discriminative features for brown dwarfs. The confusion matrices reveal that the CNN never predicts the brown dwarf class, neither for brown dwarf nor galaxy inputs. Instead, all brown dwarf cutouts are consistently misclassified as galaxies. This behaviour indicates that the model is overfitting to the dominant classes and unable to represent the minority class effectively, primarily due to class imbalance and the lack of diversity in the brown dwarf training set. Resolving this would require a substantially larger and more varied set of brown dwarf observations, along with potential use of advanced methods such as synthetic data generation, class re-weighting, or few-shot learning (Parnami & Lee, 2022).

5.3 Analysis of CNN Misclassifications

To better understand the source of the CNN's classification errors, we examine patterns in the misclassified inputs and evaluate how performance changes under restricted input conditions. One key trend among misclassified low-redshift galaxies is that they often exhibit both bright central components and elevated background flux levels. This combination appears to obscure morphological distinctions that typically differentiate low-redshift sources from their high-redshift counterparts, leading to ambiguity in the model's learned representations. In particular, when spatial contrast is low, the network seems to conflate these systems with high-redshift galaxies, especially if spectral cues are less prominent or partially suppressed.

To isolate the model's spectral dependence, we retrain and test the CNN using only the final photometric band (F444W), which lies at the red end of the observed spectrum. This forces the model to rely exclusively on the reddest available information and eliminates access to the bluer filters where Lyman break signatures are most prominent. Under these conditions, the model's classification performance drops to near-random (confusion matrices yield $\sim 50\%$ classification rates), providing evidence that the network is primarily identifying redshifted spectral breaks as its dominant discriminative feature. These findings are consistent with the results of the photometric ablation study and further highlight the model's spectral bias. While this behaviour is expected in a redshift classification task, it also exposes the CNN's limitations when spectral cues are ambiguous or masked by photometric noise.

6 Summary

This study presents a comprehensive pipeline for classifying galaxies in JWST NIRCam imaging data, combining supervised and semi-supervised machine learning approaches. Using a balanced dataset of high- and low-redshift galaxies, alongside a small number of brown dwarf contaminants, we demonstrate that CNNs are highly effective in distinguishing between galaxy populations, achieving classification accuracies of $97 - 98\%$ and robust performance even on galaxies with ambiguous redshift estimates.

Our results underscore the critical role of mid-infrared bands in redshift classification, likely due to their relatively low noise levels. Both data augmentation and spectral–spatial residual learning significantly enhance model generalisation; however, the architectural separation of spectral and spatial processing introduces limitations. The CNN frequently misclassifies brown dwarfs as galaxies, highlighting the need for more discriminative features to better disentangle these contaminants. Semi-supervised approaches, including HDBSCAN and a RF, successfully capture broad redshift structure and isolate extreme high-redshift populations, but struggle with ambiguous cases, such as sources with bimodal PDFs, indicating that additional morphological or spectral constraints may be necessary to improve classification performance.

Overall, this work establishes a strong foundation for high-redshift galaxy classification using deep learning, and demonstrates both the strengths and current limitations of CNN-based models in a realistic astrophysical setting. Future work will focus on refining the architecture to capture multi-scale features, improving brown dwarf separation, and incorporating morphological indicators to enhance performance on degenerate photometric cases.

7 Acknowledgements

Funding for this project was generously provided by Google DeepMind, the Royal Academy of Engineering, and the Hg Foundation. The CNN model developed for this study is publicly available and can be accessed via DOI: 10.5281/zenodo.16738262.

References

N. J. Adams, C. J. Conselice, et al. Epochs paper ii: The ultraviolet luminosity function from $7.5 < z < 13.5$ using 180 square arcminutes of deep, blank-fields from the pearls survey and public jwst data, 2024.

N. J. Adams, D. Austin, et al. The impact of medium-width bands on the selection, and subsequent luminosity function measurements, of high-z galaxies, 2025.

V. Asadi, H. Haghi, et al. Semi-supervised classification of stars, galaxies and quasars using k-means and random forest, 2025.

D. Austin. *Inferring the Properties of Star-Forming Galaxies in the Epoch of Reionization with JWST*. PhD thesis, University of Manchester, Manchester, UK, In Prep.

D. Austin et al. Galaxy catalogue for the upcoming epochs v2 paper. 2025, In Prep. URL `https://github.com/duncanaustin98/galfind.git`.

M. R. Beck, C. Scarlata, et al. Integrating human and machine intelligence in galaxy morphology classification tasks. *Monthly Notices of the Royal Astronomical Society*, 476(4):5516–5534, 2018. doi: 10.1093/mnras/sty464.

E. Bertin and S. Arnouts. SExtractor: Software for source extraction. *Astronomy & Astrophysics Supplement Series*, 117:393–404, 1996. doi: 10.1051/aas:1996164.

J. Bialopetravičius, D. Narbutis, et al. Deriving star cluster parameters with convolutional neural networks: I. age, mass, and size. *Astronomy & Astrophysics*, 621:A103, 2019. doi: 10.1051/0004-6361/201833833.

L. Bisigello, K. I. Caputi, et al. The impact of jwst broadband filter choice on photometric redshift estimation. *The Astrophysical Journal Supplement Series*, 227(2):19, 2016. doi: 10.3847/0067-0049/227/2/19.

G. B. Brammer, P. G. van Dokkum, et al. Eazy: A fast, public photometric redshift code. *The Astrophysical Journal*, 686(2):1503–1513, 2008.

L. Breiman. Random forests. *Machine Learning*, 45 (1):5–32, 2001. doi: 10.1023/A:1010933404324.

J. S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. F. Fogelman-Soulié and J. Hérault (eds.), *Neurocomputing: Algorithms, Architectures*, pp. 227–236. Springer Berlin Heidelberg, Berlin, Heidelberg, 1990.

C. J. Burke, P. D. Aleo, et al. Deblending and classifying astronomical sources with mask r-cnn deep learning. *Monthly Notices of the Royal Astronomical Society*, 490(3):3952–3965, 2019. doi: 10.1093/mnras/stz2845.

T. Cheng, C. J. Conselice, et al. Optimizing automatic morphological classification of galaxies with machine learning and deep learning using dark energy survey imaging. *Monthly Notices of the Royal Astronomical Society*, 493(3):4209–4228, 2020. doi: 10.1093/mnras/staa501.

T. Clausen, C. L. Steinhardt, et al. Performance of photometric template fitting for ultra-high-redshift galaxies. *Astronomy & Astrophysics*, 697:A160, 2025. doi: 10.1051/0004-6361/202453247.

C. J. Conselice, N. Adams, et al. Epochs i. the discovery and star forming properties of galaxies in the epoch of reionization at $6.5 < z < 18$ with pearls and public jwst data, 2024.

D. Dablain, K. N. Jacobson, et al. Understanding cnn fragility when learning with imbalanced data, 2022.

T. Dahlen, B. Mobasher, et al. A critical assessment of photometric redshift methods: A candels investigation. *The Astrophysical Journal*, 775(2):93, 2013. doi: 10.1088/0004-637x/775/2/93.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pp. 886–893. IEEE, 2005.

F. D'Eugenio, A. J. Cameron, et al. Jades data release 3 – nirspec/msa spectroscopy for 4,000 galaxies in the goods fields, 2024.

M. Dickinson, M. Giavalisco, et al. *The Great Observatories Origins Deep Survey*, pp. 324–331. Springer-Verlag, 2002. doi: 10.1007/10899892_78.

M. Dominique, A. N. Zhukov, et al. First detection of solar flare emission in mid-ultraviolet balmer continuum. *The Astrophysical Journal Letters*, 867(2): L24, 2018. doi: 10.3847/2041-8213/aaeace.

P. Dubath, L. Rimoldini, et al. Random forest automated supervised classification of hipparcos periodic variable stars: Classification of hipparcos periodic stars. *Monthly Notices of the Royal Astronomical Society*, 414(3):2602–2617, 2011. doi: 10.1111/j.1365-2966.2011.18575.x.

S. R. Dubey, S. K. Singh, et al. Activation functions in deep learning: A comprehensive survey and benchmark, 2022.

D. J. Eisenstein, B. Robertson, et al. The jwst advanced deep extragalactic survey (jades). *The Astrophysical Journal Supplement Series*, 265(1):1, 2023.

K. Fawagreh, M. M. Gaber, et al. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering*, 2(1):602–609, 2014. doi: 10.1080/21642583.2014.956265.

T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

A. Ferrara. The eventful life of gs-z14-0, the most distant galaxy at redshift z = 14.32. *Astronomy & Astrophysics*, 689:A310, 2024. doi: 10.1051/0004-6361/202450944.

J. Frontera-Pons, F. Sureau, et al. Representation learning for automated spectroscopic redshift estimation. *Astronomy & Astrophysics*, 625:A73, 2019. doi: 10.1051/0004-6361/201834295.

K. Fu, C. J. Conselice, et al. Dust extinction measures for $z \sim 8$ galaxies using machine learning on jwst imaging, 2024.

M. Giavalisco. Lyman-Break Galaxies. *Annual Review of Astronomy and Astrophysics*, 40:579–641, 2002. doi: 10.1146/annurev.astro.40.121301.111837.

I. Goodfellow, H. Lee, et al. Measuring invariances in deep networks. In *Advances in Neural Information Processing Systems*, volume 22, pp. 646–654. Curran Associates, Inc., 2009.

T. Harvey and D. Austin. Brown dwarf fitter github repository, 2025. URL `https://github.com/tHarvey303/BD-Finder.git`.

K. He, X. Zhang, et al. Deep residual learning for image recognition, 2015a.

K. He, X. Zhang, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015b.

B. Henghes, C. Pettitt, et al. Benchmarking and scalability of machine-learning methods for photometric redshift estimation. *Monthly Notices of the Royal Astronomical Society*, 505(4):4847–4856, 2021. doi: 10.1093/mnras/stab1513.

B. Hovis-Afflerbach, C. L. Steinhardt, et al. Identifying and repairing catastrophic errors in galaxy properties using dimensionality reduction. *The Astrophysical Journal*, 908(2):148, 2021. doi: 10.3847/1538-4357/abd329.

J. Hu, L. Shen, et al. Squeeze-and-excitation networks, 2019.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

S. Jamal, V. Le Brun, et al. Automated reliability assessment for spectroscopic redshift measurements. *Astronomy & Astrophysics*, 611:A53, 2018. doi: 10.1051/0004-6361/201731305.

T. Karalidi, M. Marley, et al. The sonora substellar atmosphere models. ii. cholla: A grid of cloud-free, solar metallicity models in chemical disequilibrium for the jwst era. *The Astrophysical Journal*, 923(2):269, 2021. doi: 10.3847/1538-4357/ac3140.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2017.

I. Kolesnikov, V. M. Sampaio, et al. Unveiling galaxy morphology through an unsupervised-supervised hybrid approach. *Monthly Notices of the Royal Astronomical Society*, 528(1):82–107, 2023. doi: 10.1093/mnras/stad3934.

A. Kuruvanthodi, D. Schaerer, et al. Strong balmer break objects at $z \sim 7 - -10$ uncovered with jwst. *Astronomy & Astrophysics*, 691:A310, 2024. doi: 10.1051/0004-6361/202451622.

I. Labbe, J. E. Greene, et al. An unambiguous agn and a balmer break in an ultraluminous little red

dot at z=4.47 from ultradeep uncover and all the little things spectroscopy, 2024.

R. L. Larson, T. A. Hutchison, et al. Spectral templates optimal for selecting galaxies at $z > 8$ with the jwst. *The Astrophysical Journal*, 958(2):141, 2023.

M. S. Marley, D. Saumon, et al. The sonora brown dwarf atmosphere and evolution models. i. model description and application to cloudless atmospheres in rainout chemical equilibrium. *The Astrophysical Journal*, 920(2):85, 2021. doi: 10.3847/1538-4357/ac141d.

M. W. McElwain, L. D. Feinberg, 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):058001, 2023. doi: 10.1088/1538-3873/acada0.

L. McInnes and J. Healy. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 33–42. IEEE, 2017. doi: 10.1109/icdmw.2017.12.

A. M. Meisner, A. C. Schneider, et al. New candidate extreme t subdwarfs from the backyard worlds: Planet 9 citizen science project. *The Astrophysical Journal*, 915(2):120, 2021. doi: 10.3847/1538-4357/ac013c.

E. Merlin, M. Castellano, et al. Euclid preparation: Xxv. the euclid morphology challenge: Towards model-fitting photometry for billions of galaxies. *Astronomy & Astrophysics*, 671:A101, March 2023. doi: 10.1051/0004-6361/202245041.

C. V. Morley, S. Mukherjee, et al. The sonora substellar atmosphere models. iii. diamondback: Atmospheric properties, spectra, and evolution for warm cloudy substellar objects, 2024.

S. Mukherjee, J. J. Fortney, et al. The sonora substellar atmosphere models. iv. elf owl: Atmospheric mixing and chemical disequilibrium with varying metallicity and c/o ratios. *The Astrophysical Journal*, 963(1):73, 2024. doi: 10.3847/1538-4357/ad18c2.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 807–814, Madison, WI, USA, 2010. Omnipress.

D. Navarro-Gironés, E. Gaztañaga, et al. The pau survey: photometric redshift estimation in deep wide fields. *Monthly Notices of the Royal Astronomical Society*, 534(2):1504–1527, 2024. doi: 10.1093/mnras/stae1686.

K. O'Shea and R. Nash. An introduction to convolutional neural networks, 2015.

M. Ouchi, Y. Ono, et al. Observations of the $lyman - \alpha$ universe. *Annual Review of Astronomy and Astrophysics*, 58(1):617–659, 2020. doi: 10.1146/annurev-astro-032620-021859.

C. Papovich, J. W. Cole, et al. Galaxies in the epoch of reionization are all bark and no bite – plenty of ionizing photons, low escape fractions, 2025.

A. Parnami and M. Lee. Learning from few examples: A summary of approaches to few-shot learning, 2022.

J. Pasquet, E. Bertin, et al. Photometric redshifts from sdss images using a convolutional neural network. *Astronomy & Astrophysics*, 621:A26, 2018.

A. Paszke, S. Gross, et al. Pytorch: An imperative style, high-performance deep learning library, 2019.

F. Pedregosa, G. Varoquaux, et al. Scikit-learn: Machine learning in python, 2018.

M. W. Phillips, P. Tremblin, et al. A new set of atmosphere and evolution models for cool t–y brown dwarfs and giant exoplanets. *Astronomy & Astrophysics*, 637:A38, 2020. doi: 10.1051/0004-6361/201937381.

D. M. W. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation, 2020.

O. Razim, S. Cavuoti, et al. Improving the reliability of photometric redshift with machine learning. *Monthly Notices of the Royal Astronomical Society*, 507(4):5034–5052, 2021. doi: 10.1093/mnras/stab2334.

M. J. Rieke, B. E. Robertson, et al. Jades initial data release for the hubble ultra deep field: Revealing the faint infrared sky with deep jwst nircam imaging, 2023.

B. E. Robertson, R. S. Ellis, et al. Early star-forming galaxies and the reionization of the universe. *Nature*, 468(7320):49–55, 2010. doi: 10.1038/nature09527.

P. Roy, S. Ghosh, et al. Effects of degradations on deep neural network architectures, 2025.

K. R. Shahapure et al. Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 747–748, 2020. doi: 10.1109/DSAA49011.2020.00096.

T. Signor, G. Rodighiero, et al. Euclid: Identifying the reddest high-redshift galaxies in the euclid deep fields with gradient-boosted trees. *Astronomy & Astrophysics*, 685:A127, 2024. doi: 10.1051/0004-6361/202348737.

S. A. Solla, E. Levin, et al. Accelerated learning in layered neural networks. *Complex Systems*, 2:625–640, 1988.

N. Srivastava, G.Hinton, et al. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15 (56):1929–1958, 2014.

S. Tanigawa, K. Glazebrook, et al. Hayate: Photometric redshift estimation by hybridising machine learning with template fitting, 2024.

J. Taran. Convolutional neural network for lyman break galaxies classification and redshift regression in desi (dark energy spectroscopic instrument), 2024.

C. Tohill, S. Bamford, et al. Exploring the morphologies of high redshift galaxies with machine learning, 2023.

C. Tohill, S. P. Bamford, et al. A robust study of high-redshift galaxies: Unsupervised machine learning for characterizing morphology with jwst up to $z \sim 8$. *The Astrophysical Journal*, 962(2): 164, 2024. doi: 10.3847/1538-4357/ad17b8.

Z. Tu, S. Wang, et al. Three brown dwarfs masquerading as high-redshift galaxies in jwst observations, 2025.

L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

B. Margalef-Bentaboland L. Wang et al. Galaxy merger challenge: A comparison study between machine learning-based detection methods. *Astronomy & Astrophysics*, 687:A24, 2024.

S. M. Wilkins, C. C. Lovell, et al. First light and reionisation epoch simulations (flares) xiv: The balmer/4000 å breaks of distant galaxies, 2023.

D. Xu and Y. Zhu. Surveying image segmentation approaches in astronomy, 2024.

B. Yue, A. Ferrara, et al. The contribution of high-redshift galaxies to the near-infrared background. *Monthly Notices of the Royal Astronomical Society*, 431(1):383–393, 2013. doi: 10.1093/mnras/stt174.

N. R. Zhong, F.and Napolitano et al. Galaxy spectra neural network (gasnet). ii. using deep learning for spectral classification and redshift predictions. *Monthly Notices of the Royal Astronomical Society*, 532(1):643–665, 2024. doi: 10.1093/mnras/stae1461.

X. Zhou, Y. Gong, et al. Spectroscopic and photometric redshift estimation by neural networks for the china space station optical survey (css-os). *The Astrophysical Journal*, 909(1):53, 2021. doi: 10.3847/1538-4357/abda3e.

A Example SEDs

Figure 7 presents example spectral energy distributions (SEDs) for galaxies at redshifts ~ 6.5 and ~ 9.0. These illustrate the characteristic shift of the Lyman break from $\sim 0.9\mu$m to $\sim 1.2\mu$m, demonstrating how this spectral feature moves through different NIRCam filters with increasing redshift.

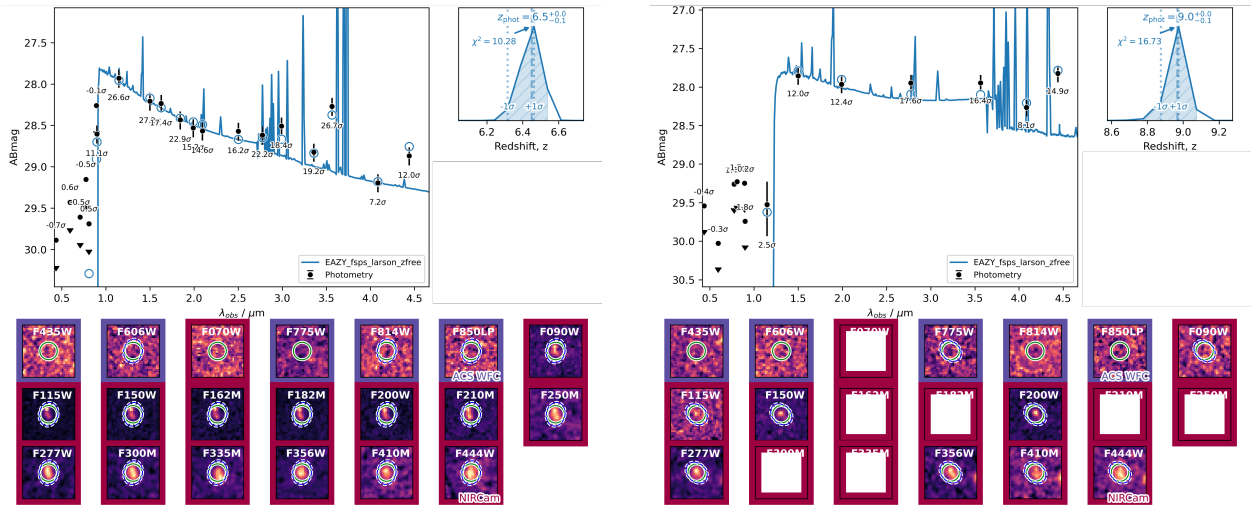


Figure 7: Galaxy SEDs at redshifts ~ 6.5 and ~ 9.0, illustrating the Lyman break shifting from $\sim 0.9\,\mu$m to $\sim 1.2\,\mu$m. NIRCam filter cutouts demonstrate how this shift across bands facilitates photometric redshift estimation. White boxes indicate missing filter coverage.

B Visualisation of CNN Architecture Components

To provide a clearer understanding of the convolutional neural network (CNN) architecture, Figures 8–10 provide visual representations of the key CNN components used in the architecture. Figure 8 shows the `SpectralResidualBlock`, while Figure 9 illustrates the `SpatialResidualBlock`. The `ChannelAttention` mechanism is shown in Figure 10.

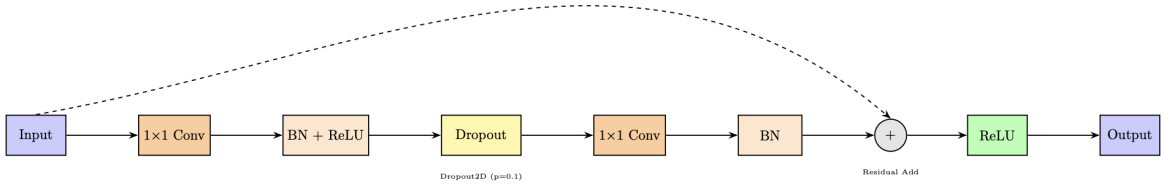


Figure 8: Diagram of a `SpectralResidualBlock`. The block consists of two sequential 1×1 convolutional layers, each followed by `BATCHNORMALISATION` and `ReLU` activation, with dropout applied after the first layer. A residual (skip) connection adds the input to the output before a final `ReLU` activation.

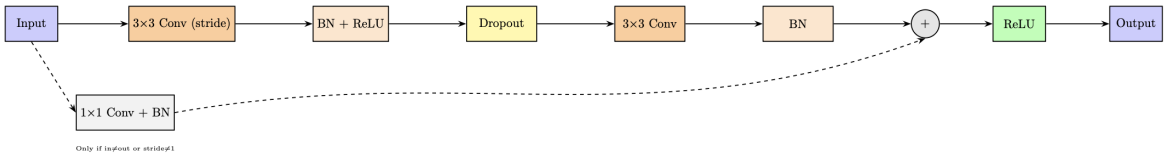


Figure 9: Diagram of a `SpatialResidualBlock`. The main path consists of two 3×3 convolutional layers, each followed by `BATCHNORMALISATION` and `ReLU` activation, with dropout applied after the first. An optional skip connection uses a 1×1 convolution and `BATCHNORMALISATION` to match dimensions when required. The output combines both paths with a final `ReLU` activation.

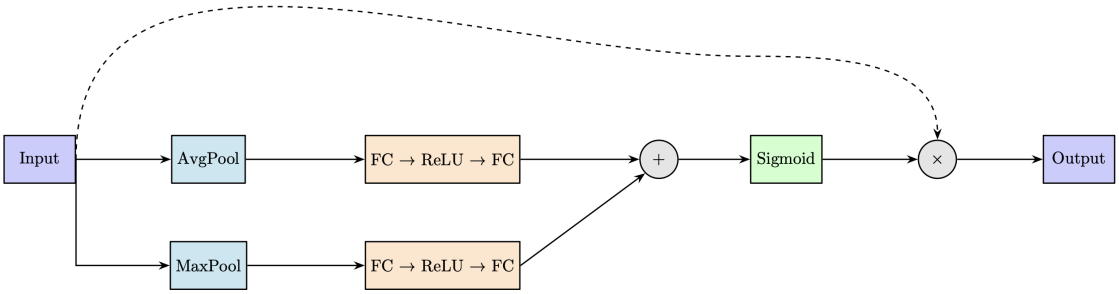


Figure 10: Diagram of the `ChannelAttention` block. Shared fully connected layers process both average and max pooled features, and the resulting outputs are combined to generate an attention map. This map is applied to re-weight the input feature channels.

C Example PDFs

To illustrate the complexity of photometric redshift estimation in ambiguous cases, Figure 11 presents the redshift probability distribution functions (PDFs) for two representative near-miss galaxies. These examples underscore the importance of examining the full redshift PDF when assessing high-redshift candidates, particularly in the presence of bimodal distributions and potential low-redshift contamination.

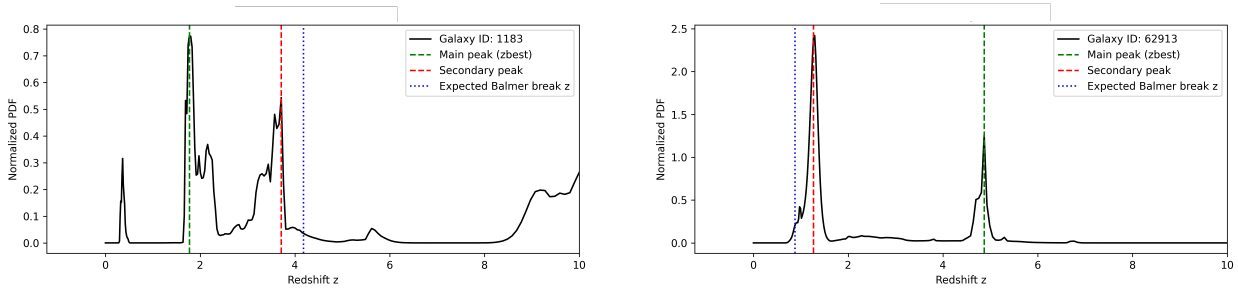


Figure 11: Photometric redshift PDFs for two near-miss cases: Galaxy 1183 (left) and Galaxy 62913 (right). Galaxy 1183 exhibits a secondary peak at a redshift lower than the best-fit value, while Galaxy 62913 shows a secondary peak at a higher redshift.

D Classification Results at Selected Confidence Thresholds

Figure 12 presents confusion matrices from two example runs applying a confidence threshold $p > 0.8$ to the CNN classification outputs. These matrices highlight the trade-off between confident predictions and the number of withheld (ambiguous) labels.

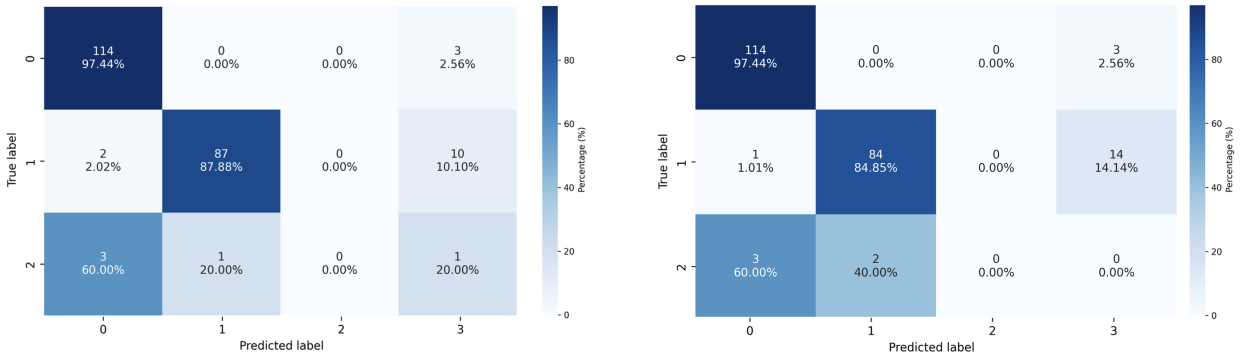


Figure 12: Confusion matrices illustrating the classification performance of the CNN outputs after applying a confidence threshold of $p > 0.8$. Class 0 represents low-redshift galaxies, class 1 corresponds to high-redshift galaxies, class 2 denotes brown dwarfs, and class 3 includes ambiguous objects.