Machine Learning for Parameters of High-Redshift Galaxies for Future JWST-NIR Cam Observations

Gunmethus Cherdchoochavalit[1,2*], Krittapas Chanchaiworawit[2], Suwicha Wannawichian[1], and Siramas Komonjinda[1]

[1] Department of Physics, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand
[2] National Astronomical Research Institute of Thailand, Chiang Mai 50180, Thailand
* gunmethus.cherdchoochavalit@gmail.com

Abstract

High-redshift galaxies provide essential insight into the early universe. However, classifying their types and estimating their properties remain challenging due to observational limitations. This study employs Principal Component Analysis (PCA) with eigen spectra fitting to decompose spectral data from galaxies in the Sloan Digital Sky Survey (SDSS). The extracted eigenvalues and eigenvectors are used as the foundation for simulating synthetic spectra that resemble observations expected from the James Webb Space Telescope (JWST). Using these simulated datasets, we develop a transformation framework to project SDSS photometric and spectroscopic data into the JWST-NIR Cam observational system. The transformed datasets are then used to train machine learning models, including Random Forest (RF), to predict galaxy redshifts and classify their types under conditions simulating JWST observations. By incorporating spectral energy distributions (SEDs), photometric data, and morphological parameters, a robust training set is created for supervised learning. Preliminary results demonstrate that the trained models can accurately predict redshifts. The research contributes to understanding the evolution of the galaxy and will support the analysis of upcoming large-scale surveys like those from JWST.

Introduction

The James Webb Space Telescope (JWST) enables observations of faint, distant galaxies at high redshift, providing crucial insight into early star formation and galaxy growth. While photometric data allows redshift estimation from color and brightness, these measurements remain uncertain compared with spectroscopic redshifts, which are costly and limited. Current JWST catalogs are incomplete due to restricted observing time, small survey coverage, low flux from faint sources, and calibration delays, resulting in selection biases that hinder studies of galaxy evolution (Clausen et al. 2024) [1].

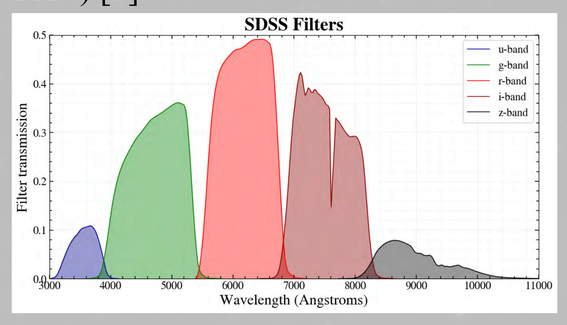


Figure 1: SDSS Filters (u, g, r, i, and z)

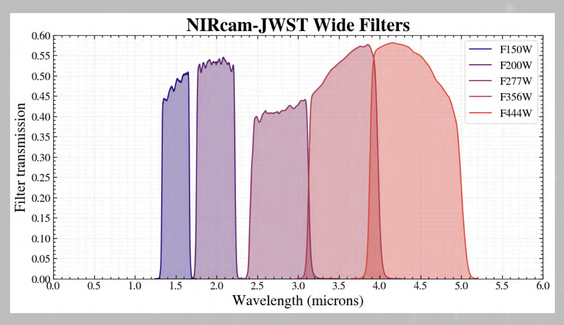


Figure 2: JWST-NIR Cam Wide Filters

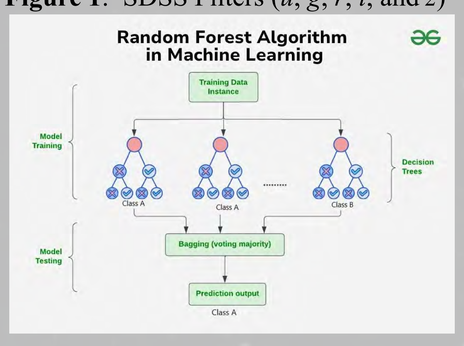


Figure 3: Random Forest Algorithm in Machine Learning

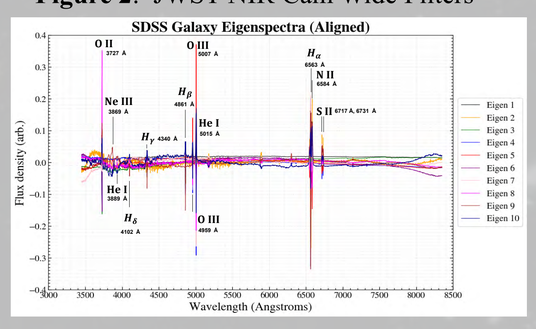


Figure 4: 10 Eigen spectra for SDSS galaxies

Machine learning offers a solution by predicting spectroscopic properties from photometric inputs. Previous work has shown improved redshift accuracy using Random Forests, Neural Networks, and Gaussian Processes over traditional SED fitting (Dahlen et al. 2013 [2]; Tanaka et al. 2018 [3]). However, JWST's unique resolution and depth complicate direct application of existing models. Leveraging legacy datasets such as SDSS, which provide extensive spectroscopic and photometric coverage, enables the development of transferable models to bridge gaps in JWST spectroscopy.

Results from JADES Catalog & CIGALE

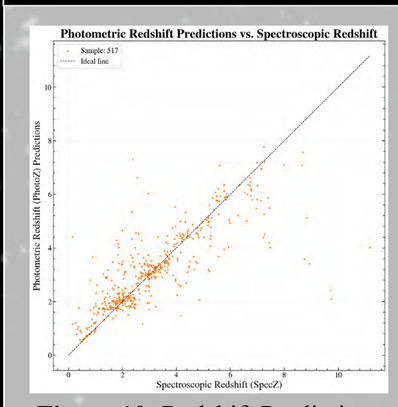


Figure 10: Redshift Prediction

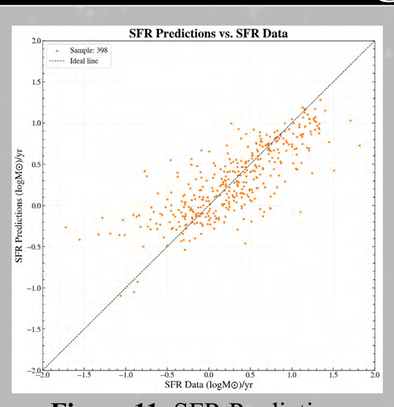


Figure 11: SFR Prediction

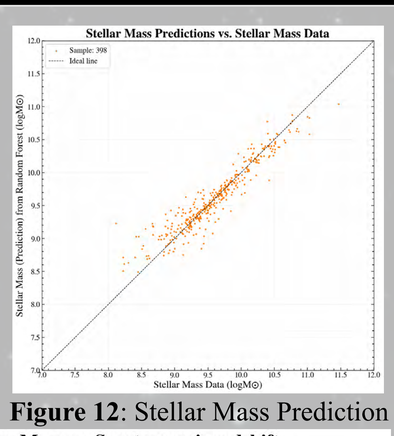


Figure 12: Stellar Mass Prediction

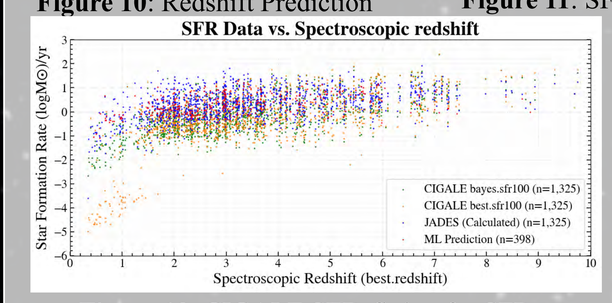


Figure 13: CIGALE SED-fitting for SFR

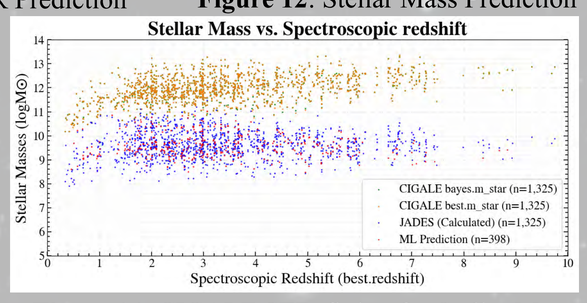


Figure 14: CIGALE SED-fitting for Stellar Mass

Methodology

Extraction Parameters

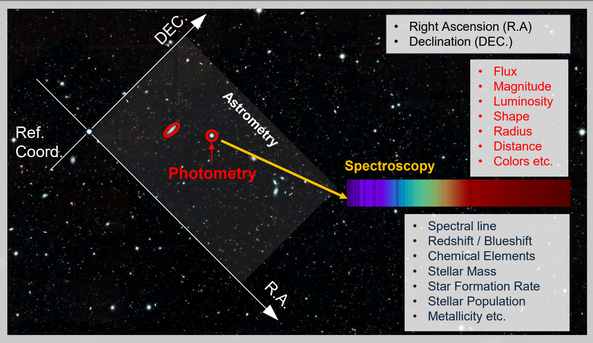


Figure 5: Extraction Photometric and Spectroscopic data

SDSS Catalog

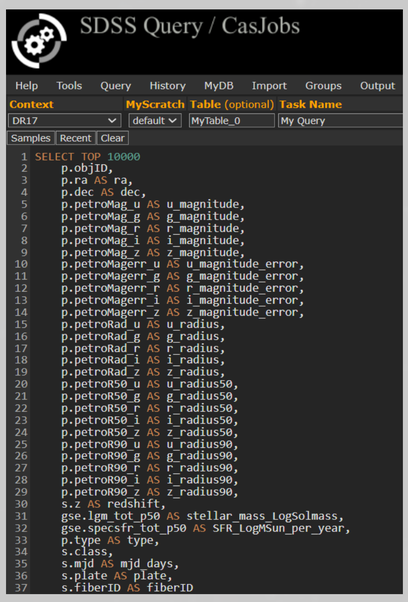


Figure 6: SQL Query for SDSS Catalog Data

PCA Decomposition [4]

$$f_R(\lambda) = \sum_{\kappa=1}^{N} a_\kappa e_\kappa(\lambda)$$

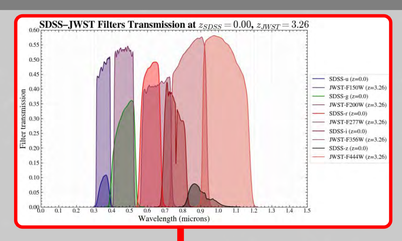
22,813 Galaxies

$e_\kappa(\lambda)$: Eigenvectors (galaxy spectral templates)
a_κ: Eigenvalue

Simulate JWST Spectrum

10 Eigenvalues + 10 Errors

Conditions: (each galaxy)
Eigen_1 > 3 × Error_1
|Eigen_2| > Error_2
SNR_u > 10
SNR_g > 20
SNR_r > 30
SNR_i > 20
SNR_z > 10



Filters Analysis

Qualified Eigenvalues (~ 11,150 galaxies)

Simulate 9 more Eigenvalues

Simulated JWST Spectrum (~ 111,500 data)

Figure 7: Transform framework from SDSS to JWST NIR-Cam System

Machine Learning for JWST Data

Simulated JWST Spectrum (~ 111,500 data)

Conditions: No Missing Data

Extract Photometric Data

Photometric JWST Data (~ 111,500 data)

70 % → Training Dataset (~ 78,050 data) → Machine Learning (Random Forest) → Model → Performance

30 % → Testing Dataset (~ 33,450 data) → Model

Results & Conclusion

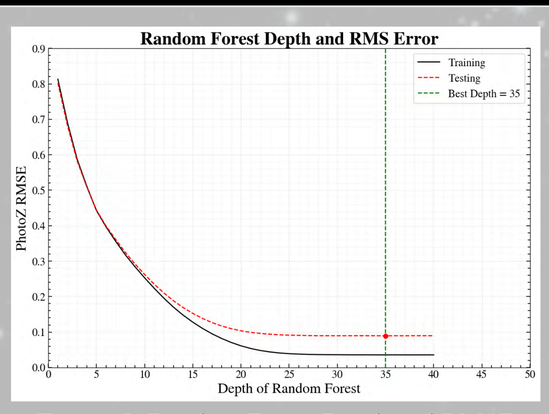


Figure 8: Random Forest Depth and RMSE

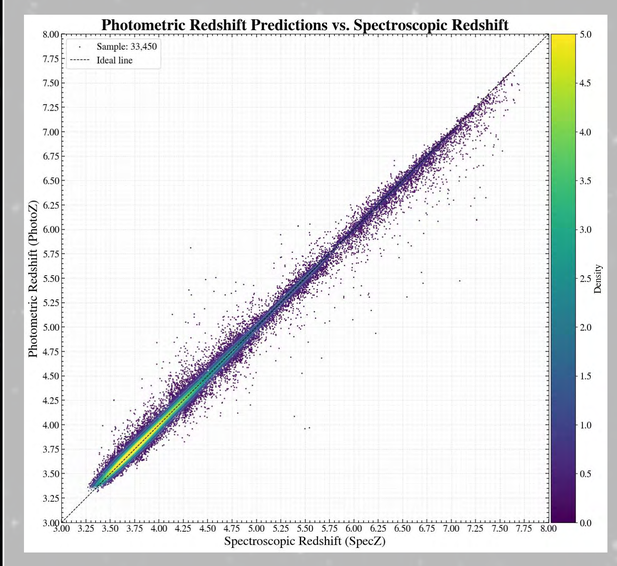


Figure 9: Prediction Redshift vs. Redshift Data

The RMS error decreases with increasing depth and flattens beyond 25 as the model reaches diminishing returns.

Best depth = 35, giving the lowest test RMS error before overfitting.

- The Random Forest model shows strong performance across the core range, with **predictions closely following the 1:1 relation** between $z \approx$ **3.2 - 5.0.**
- The high-density region indicates accurate and stable photo-z estimates with minimal scatter.
- At $z >$ **5**, prediction scatter increases and the model tends to **slightly under-estimate redshifts**.
- Some high-z objects are **shifted toward lower predicted z**, reflecting reduced sensitivity in this regime.
- Overall, **the model maintains good redshift** ordering and demonstrates reliable performance across the full redshift range.

References

[1] Clausen, T., et al. (2024). Performance of photometric template fitting for ultra-high redshift galaxies. Astronomy & Astrophysics, 671, A1.

[2] Dahlen, T., et al. (2013). A Critical Assessment of Photometric Redshift Methods: A CANDELS Investigation. The Astrophysical Journal, 775(2), 93.

[3] Tanaka, M., et al. (2018). Photometric redshifts for the Hyper Suprime-Cam Subaru Strategic Program. Publications of the Astronomical Society of Japan, 70(SP1).

[4] C. W. e. a. Yip. Principal component analysis of SDSS galaxies. The Astronomical Journal,128:585 and 2603, 2004.