Machine Learning Approaches for Predicting Redshift and Star Formation Rates from Multi-Modal Astronomical Data

February 19, 2025

Abstract

In this work, we explore a variety of machine learning (ML) methods applied to the estimation of redshift (Z) and star formation rate (SFR) from multi-modal astronomical data drawn from the Sloan Digital Sky Survey (SDSS). Our approach combines both photometric images and spectroscopic data. We experiment with classical regression techniques, convolutional neural networks (CNNs), Bayesian CNNs integrated with active learning, and dimensionality reduction tools such as t-SNE and UMAP. In particular, we compare models trained on individual modalities (images only, spectra only) with those that integrate the complementary information from both sources. The results show that a fusion strategy—combining a model based on photometry with one based on spectroscopy—yields improved accuracy in predicting redshift and the statistical SFR parameters (AVG, ENTROPY, MEDIAN, MODE, P16, P2P5, P84, P97P5).

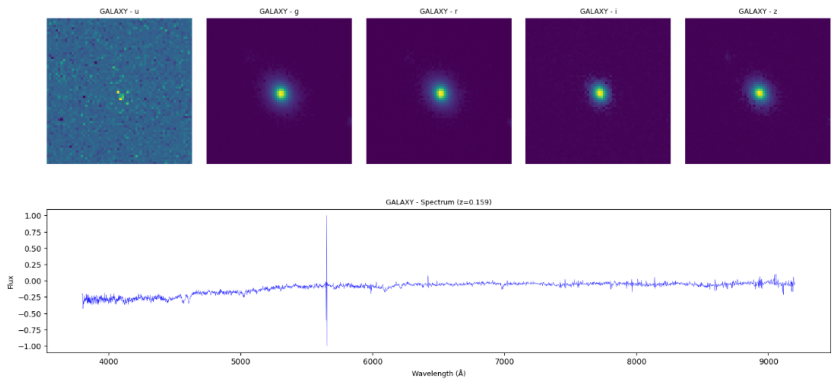

Figure 1: Overview of the multi-modal data fusion architecture.

1 Introduction

Accurate measurements of astrophysical parameters such as redshift (Z) and star formation rate (SFR) are essential for understanding galaxy evolution and cosmology. Traditionally, redshift is determined via spectroscopic observations and SFR through a variety of photometric and spectral indicators. With the advent of large-scale surveys like the SDSS, large volumes of both photometric images and spectral data are available. This has paved the way for applying advanced machine learning (ML) techniques to combine these heterogeneous data types in order to improve predictive accuracy.

Recent developments in ML have shown that multi-modal approaches—where information from different data sources is integrated—can overcome limitations inherent in using a single modality [8]. In astrophysics, such multi-modal techniques can merge the spatial and morphological information from photometry with the detailed line features present in spectra. Our work focuses on leveraging this complementarity by implementing two parallel models (one for images and one for spectra) whose outputs are later fused to predict redshift and SFR.

In the SDSS dataset, SFR is represented by several columns, including:

- **AVG**: Mean value of the SFR distribution,
- **ENTROPY**: Entropy (a measure of uncertainty),

- **MEDIAN**: Median SFR value,
- **MODE**: Most probable SFR,
- **P16**: 16th percentile,
- **P2P5**: 2.5th percentile,
- **P84**: 84th percentile,
- **P97P5**: 97.5th percentile.

The use of these multiple statistical measures allows our model to capture both the central tendency and the spread of SFR estimates.

2 Multi-Modal Data in Astrophysics

Astronomical data are naturally multi-modal. Photometric images provide key insights into the morphology and global features of galaxies. In contrast, spectroscopic data offer detailed information on the chemical composition, kinematics, and redshift of these objects.

Combining these modalities is especially beneficial for tasks such as predicting redshift and SFR. For instance, while photometry captures overall brightness and color gradients, spectroscopy reveals emission and absorption line features that are critical for precise redshift estimation. By fusing the two, a machine learning model can leverage the strengths of both data types.

In our approach, separate preprocessing pipelines are used for each modality. Photometric images are normalized and augmented to improve generalization, while spectral data are carefully calibrated and feature–extracted to highlight relevant spectral lines. The resulting feature sets serve as inputs to our ML models.

3 Related Work and Literature Review

A number of studies have addressed the prediction of SFR and redshift using different combinations of photometric and spectral data.

In [1], the authors demonstrated that combining redshift and photometric information significantly improved SFR predictions compared to traditional methods. Similarly, [2] focused on SFR prediction using only photometric data, highlighting the potential of deep learning techniques in handling large image datasets.

The integration of additional data sources has also been shown to be advantageous. For example, [3] combined redshift with WISE (Wide-field Infrared Survey Explorer) data to predict SFR, underscoring the role of infrared information. In [4], the authors presented a method to predict the specific star formation rate (sSFR) using a combination of photo-Z, spectrum-Z, and photometric measurements, emphasizing the complementary nature of these modalities.

Other works have extended the approach further. In [5], SFR and stellar mass (M^*) were predicted by integrating photometry with emission line data, metallicity, and dust indices (such as D4000). In [6], a similar multi-modal framework was used to combine photometry, redshift, and WISE data to yield improved estimates of SFR and M^*.

More recent advancements include the development of Bayesian CNNs and active learning techniques. In [7], a framework combining Bayesian convolutional neural networks with active learning and data visualization tools (e.g., t-SNE, UMAP) was proposed to enhance training efficiency and uncertainty quantification. Furthermore, [8] discusses how modern AI models can handle heterogeneous data formats, allowing the creation of universal systems capable of processing diverse astronomical data.

Finally, the challenge of modality imbalance—where the model may rely too heavily on one data source—is addressed in [9]. This study provides strategies to ensure that each modality contributes equitably to the prediction, a topic we also explore in our work.

4 Methodology

The primary objective of our methodology is to construct and evaluate ML models that effectively combine information from photometric images and spectroscopic data to predict redshift and SFR.

4.1 Model Selection and Architecture

Based on the reviewed literature [1–6], we selected a diverse set of models ranging from traditional regression techniques to deep learning architectures. Our baseline includes:

- **Classical Regression Models:** Linear regression and Random Forest regressors for initial benchmarks.
- **Convolutional Neural Networks (CNNs):** For processing photometric images, taking advantage of their spatial structure.
- **Fully-Connected Neural Networks:** For spectral data where sequential features are important.

To harness the complementary strengths of each modality, we designed a dual-stream network architecture. In this framework, one branch is dedicated to processing images while the other focuses on spectra. Their latent representations are then merged via a fusion layer (using concatenation or more advanced ensemble techniques) to produce final predictions for Z and SFR.

4.2 Training Procedure and Evaluation Metrics

All experiments were implemented in a Jupyter Notebook environment to facilitate reproducibility and iterative testing. Data preprocessing included normalization, augmentation (for images), and careful feature extraction (for spectra).

We evaluated model performance using several metrics:

- **Root Mean Squared Error (RMSE):** Quantifies the deviation of predictions from true values.
- **Coefficient of Determination ($\mathbf{R}^2$):** Measures the proportion of variance explained by the model.
- **Mean Absolute Error (MAE):** Offers an interpretable measure of average prediction error.

Hyperparameter tuning was performed via cross-validation, and particular care was taken to balance the learning from each modality. This is important to avoid the issue of modality imbalance discussed in [9].

5 Experimental Setup

Our experimental framework was designed to compare three primary configurations:

1. **Spectral-Only Models:** Trained exclusively on spectroscopic data.
2. **Image-Only Models:** Trained exclusively on photometric images.
3. **Combined Multi-Modal Models:** Employing fusion techniques to integrate outputs from both the spectral and image branches.

For the spectral-only experiments, each object's spectrum is represented as a numpy.ndarray with 4620 columns of float values, capturing detailed spectral features. In the image-only experiments, the photometric data for each object consists of 5 images of size 64×64 pixels, corresponding to the SDSS filters: u, g, r, i, and z. The multi-modal models were implemented using both early fusion (combining the raw image and spectral features) and late fusion (combining the outputs of independently trained sub-models for the two modalities). The training and evaluation phases were standardized across all experi- ments. We maintained the same train-validation-test splits and recorded the execution time and prediction precision for each configuration. The overall goal was to assess not only the accuracy improvements from multi-modal integration but also the computational trade-offs involved.

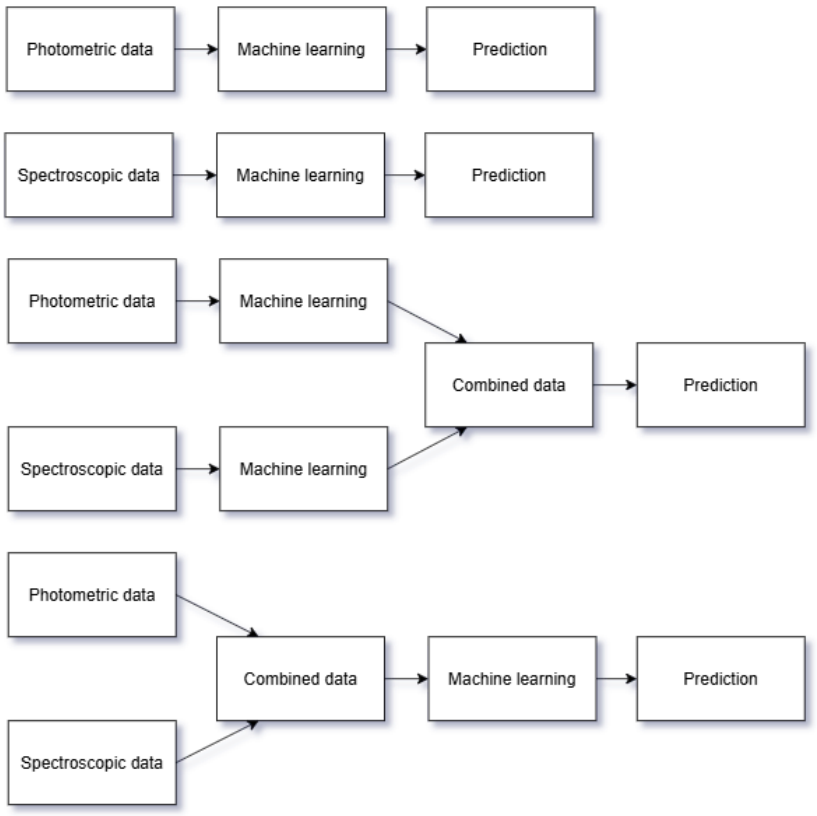

Figure 2: Ways to combine photometric and spectroscopic data for machine learning.

References

[1] Author A. et al. (Year). *Predicting SFR from Redshift and Photometry. Monthly Notices of the Royal Astronomical Society*, 493(4), 4808. https://academic.oup.com/mnras/article/493/4/4808/5758312

[2] Author B. et al. (Year). *SFR Prediction from Photometry Using Deep Learning Techniques. Monthly Notices of the Royal Astronomical Society*, 486(1), 1377. https://academic.oup.com/mnras/article/486/1/1377/5420450

[3] Author C. et al. (Year). *Integrating Redshift and WISE Data for SFR Estimation. Astronomy & Astrophysics.* https://www.aanda.org/articles/aa/abs/2019/02/aa33972-18/aa33972-18.html

[4] Author D. et al. (Year). *Predicting Specific Star Formation Rate Using Multi–Modal Data. Monthly Notices of the Royal Astronomical Society.* https://doi.org/10.1093/mnras/stw2476

[5] Author E. et al. (Year). *SFR and Stellar Mass Predictions from Photometry, Emission Lines, Metallicity, and Dust Indicators. Monthly Notices of the Royal Astronomical Society.* https://doi.org/10.1111/j.1365-2966.2004.07881.x

[6] Author F. et al. (Year). *Multi–Modal Approaches for SFR and Stellar Mass Estimation. The Astrophysical Journal.* https://doi.org/10.1088/0067-0049/219/1/8

[7] Author G. et al. (Year). *Improving Neural Network Training in Astronomy Using Bayesian CNNs and Active Learning. Proceedings of the ASP Conference Series.* https://articles.adsabs.harvard.edu/pdf/2024ASPC..535...91P

[8] Author H. et al. (Year). *Modern AI Models for Multi–Modal Data Integration in Astronomy. arXiv preprint arXiv:2411.06284v2.* https://arxiv.org/abs/2411.06284v2

[9] Author I. et al. (Year). *Addressing Modality Imbalance in Multi–Modal Neural Networks. Proceedings of the International Conference on Machine Learning.* `https://proceedings.mlr.press/v162/wu22d.html`