Bachelor's thesis

# APPLICATION OF MACHINE LEARNING TO PREDICT STAR FORMATION RATES IN SDSS DATA

**Bc. Farukh Rustamov**

Faculty of Information Technology
Department of Applied Mathematics
Supervisor: ————
April 29, 2025

FACULTY OF INFORMATION TECHNOLOGY CTU IN PRAGUE

# Assignment of bachelor's thesis

| | |
|---|---|
| **Title:** | Experiment with Machine Learning on Hierarchical Multi-Modal Astronomical Data |
| **Student:** | Farukh Rustamov |
| **Supervisor:** | RNDr. Petr Škoda, CSc. |
| **Study program:** | Informatics |
| **Branch / specialization:** | Artificial Intelligence 2021 |
| **Department:** | Department of Applied Mathematics |
| **Validity:** | until the end of summer semester 2025/2026 |

## Instructions

Current astronomy is flooded by Petabyte-scaled data detected in all frequencies of the electromagnetic spectrum. In order to find new physically interesting objects and phenomena, advanced machine learning of such data becomes a natural part of data analysis. One of the most important astronomical surveys is the Sloan Digital Sky Survey (SDSS) containing several millions of sky images in five spectral filters and a similar amount of spectra observed by the same telescope. It gives a unique opportunity to study advanced machine learning methods applied to multi-dimensional and dimensionally multi-modal data. A combination of SDSS multi-color images and spectra exposed at different times results in a multi-dimensional semi-sparse datacube of about a hundred terabytes in size. For this purpose there was recently developed a parallel processing and storage framework Hierarchical Semi-Sparse Cubes (HiSS -Cube). HiSS-Cube also handles the uncertainty estimates and pre-computes the data in several scales, allowing fast interactive zooming of a given part of the sky and quick machine learning experiments on coarse data in order to identify the interesting parts of latent space before focusing on them in a higher resolution.

A unique HiSS-Cube design allows interesting experiments with multi-modal and hierarchically structured multi-scale data.

The main tasks are:

---

*Electronically approved by Ing. Magda Friedjungová, Ph.D. on 10 November 2024 in Prague.*

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

1) Install the HiSS-Cube system and download the data required for its run (SDSS images and spectra of some selected parts of the sky)
2) Identify interesting science cases where the machine learning methods trained on a combination of multi-modal data (i.e. images and spectra treated together) are expected to give better accuracy against the combination of results of methods trained on each type of modality separately.
3) Perform experiments with different ML methods (e.g. classification, regression, clustering, tSNE, CNN) on several data samples and analyze results. Compare the performance on combined multi-modal data with single-modal experiments.
4 ) Use HiSS-Cube to get all pre-computed resolutions (i.e. images and spectra of different sizes with various degrees of smearing) of the same sky region.
5) Perform simple experiments (e.g. star-galaxy-classification) on different scales of the same data and compare execution time concerning the precision.
6) (optional) Try to get access to the large cluster and perform the experiments on the whole SDSS archive

The recommended literature will be delivered by the supervisor of the thesis.

*Electronically approved by Ing. Magda Friedjungová, Ph.D. on 10 November 2024 in Prague.*

*I would like to express my sincere gratitude to my supervisor,* ***RNDr. Petr Škoda, CSc.****, for his valuable guidance, insightful feedback, and continuous support throughout the development of this thesis.*
*I would also like to thank* ***Ing. Ondřej Podsztavek*** *for his expert advice and assistance with machine learning methods, which significantly contributed to the quality and depth of the experimental work.*
*All computations for this thesis were carried out on the RCI cluster, providing access to high-performance computing resources and enabling more complex and large-scale machine learning experiments. The authors acknowledge the support of the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics".*
*The access to the computational infrastructure of the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics" is also gratefully acknowledged. Most of the experiments and data processing were carried out using the RCI cluster.*
*We further acknowledge the Sloan Digital Sky Survey (SDSS) [1], whose publicly released photometric and spectroscopic data formed the foundation of our analysis—without SDSS, the work presented here would not have been possible.*
*Finally, we thank the HiSS-Cube pipeline and its creator, Ing. Jiří Nadvorník, Ph.D. [2], for processing the SDSS data into a scalable, multi-resolution semi-sparse data cube that preserves measurement uncertainties and makes interactive visualization and machine-learning experiments on large astronomical datasets straightforward.*

# Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis. I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular that the Czech Technical University in Prague has the right to conclude a license agreement on the utilization of this thesis as a school work under the provisions of Article 60 (1) of the Act.

In Prague on April 29, 2025

## Abstract

In this thesis, we investigate the application of machine learning methods to predict the star formation rate (SFR) in astronomical objects based on photometric and spectroscopic data from the Sloan Digital Sky Survey (SDSS).



**Keywords** machine learning, SDSS, star formation rate, spectroscopy, photometry



## Abstrakt

V této diplomové práci zkoumáme použití metod strojového učení pro predikci rychlosti formování hvězd (SFR) v astronomických objektech na základě fotometrických a spektroskopických dat ze Sloan Digital Sky Survey (SDSS).



**Klíčová slova** strojové učení, SDSS, rychlost formování hvězd, spektroskopie, fotometrie.

# Contents

# List of Figures

## List of Tables



## List of code listings

# List of abbreviations

| | |
|---|---|
| SDSS | Sloan Digital Sky Survey |
| SFR | Star Formation Rate |
| CNN | Convolutional Neural Network |
| MFCC | Mel-Frequency Cepstral Coefficients |
| MAE | Mean Absolute Error |
| RMSE | Root Mean Square Error |
| NMAD | Normalized Median Absolute Deviation |
| DT | Decision Tree |
| VGG | Visual Geometry Group |
| ML | Machine Learning |
| HDF5 | Hierarchical Data Format version 5 |
| RCI | Research Computing Infrastructure |
| MLP | Multilayer Perceptron |

# Chapter 1

# Introduction

## 1.1 General Description and Relevance of the Study

In recent years, multimodal machine learning has become a rapidly advancing area of research with applications ranging from autonomous driving and medical diagnostics to astronomical data analysis. The integration of different data types—such as images, text, audio, and structured signals—enables models to capture richer representations and make more accurate predictions in complex domains.

In astrophysics, large-scale surveys like the Sloan Digital Sky Survey (SDSS) [3] provide both photometric and spectroscopic data for millions of celestial objects. These complementary modalities offer unique views: images capture structural and morphological features, while spectra encode detailed physical and chemical properties.

This thesis investigates the application of multimodal machine learning techniques to predict the **star formation rate (SFR)** [4] in galaxies using data from SDSS. The motivation lies in the need to efficiently process massive astronomical datasets and build models that leverage the strengths of both image-based and spectroscopic inputs.

## 1.2 SDSS Data Description

The SDSS dataset provides a unique opportunity to study the properties of astronomical objects using comprehensive observations. Each object in the sample is characterized by the following components:

- **Five-Band Photometry.** For each object, five images are available corresponding to different spectral bands (denoted as $u$, $g$, $r$, $i$, and $z$) [5]. Each image captures a specific portion of the spectrum, enabling a detailed analysis of the structural and physical properties of the objects.

- **Spectroscopic Data.** In addition to the photometric images, each object is provided with a spectrum that offers information on its chemical composition, temperature, and dynamics.

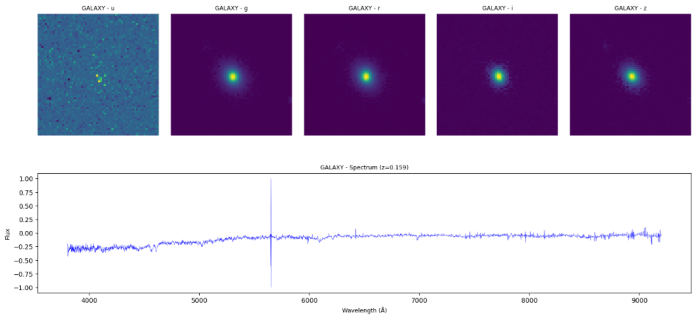



**Figure 1.1** An example of an object. Top 5 pixel photos, bottom a spectrum.

## 1.3 SDSS Data Releases

The Sloan Digital Sky Survey issues a sequence of incremental Data Releases (DR1, DR2, ...), each reprocessing the full imaging and spectroscopic dataset through updated reduction pipelines and adding newly acquired observations. The original technical summary of SDSS is given by York et al. [3], and DR7 represents the completion of the Legacy Survey, covering over 8000 deg$^2$ with more than 1.6 million galaxy spectra [6]. Subsequent releases under SDSS-III and SDSS-IV (e.g., DR13, DR14) expanded the footprint, incorporated the BOSS and eBOSS redshift programs, and further improved photometric calibration and spectrograph performance [7].

In this thesis we primarily use data from SDSS Data Release 7 (DR7) [1]. Each subsequent release extends sky coverage, improves calibration of photometry and spectroscopy, and adds new object classifications. Choosing the appropriate release is crucial, since it directly impacts the depth and quality of our SFR predictions.

## 1.4 Prediction of Star Formation Rate (SFR)

In this work, we treat the star formation rate (SFR) as a continuous regression target. SFR quantifies the mass of gas converted into stars per year in a galaxy, measured in $M_\odot\,\mathrm{yr}^{-1}$, and plays a central role in galaxy evolution studies [8]. In the SDSS catalog, we use the field `AVG`, which gives the mean of the posterior distribution of $\log_{10}(\mathrm{SFR})$ for each object [9]. Our objective is to learn a mapping

$$(\,\text{images, spectra}\,) \longrightarrow \log_{10}(\mathrm{SFR})$$

using various machine learning models.

### 1.4.1 Prediction Experiments

To assess the value of each data modality, we perform three sets of experiments:

- **Photometry-only.** Train and evaluate models using only the $u, g, r, i, z$ image cutouts.
- **Spectroscopy-only.** Train and evaluate models using only the one-dimensional spectra.
- **Multimodal fusion.** Combine image and spectral features via both early-fusion (feature concatenation) and late-fusion (prediction averaging) strategies.

### 1.4.2 Role of Spectroscopy vs. Photometry

Spectroscopic data provide direct physical diagnostics—emission-line luminosities (e.g., H$\alpha$) which scale with instantaneous SFR, as well as redshift measurements for distance correction [8]. Photometric images encode morphological details, color gradients, and integrated broadband flux, reflecting the galaxy's stellar population and dust content. By fusing these complementary views, our models can leverage both fine-scale spectral physics and global structural cues, leading to more robust and accurate SFR predictions.

## 1.5 Research Challenges

Working with the SDSS data presents several challenges: Working with the SDSS data presents several challenges:

1. **Data Filtering.** The SDSS SFR catalog originally contains over 4.8 million entries, but only a fraction have both reliable multi-band cutouts and valid SFR measurements. We must exclude objects with missing photometry or spectroscopy, undefined SFR values (NaN or the placeholder $-99$), and non-galactic sources, reducing the sample to a few $\times 10^4$ galaxies suitable for regression [10].
2. **Quality of Images and Spectra.** The HiSS-Cube pipeline provides four image resolutions (64×64 to 8×8 px) and four spectral samplings (4620 to 577 bins). While higher resolutions capture finer morphological and spectral features, they also incur substantially greater computational cost and risk overfitting; lower resolutions run faster but may smooth out diagnostically important details. Striking the optimal balance is non-trivial [2].
3. **Multiple Objects in One Image.** SDSS cutouts sometimes include overlapping galaxies or stars, leading to blended light profiles that confuse

downstream feature extractors. To ensure each input represents a single target galaxy, we apply automatic segmentation via thresholding and connected-component labeling, flagging and removing multi-object cutouts [11, 12].

## 1.6 Objectives and Tasks

The primary objective of this thesis is to develop an optimal methodology for predicting SFR using SDSS data. To achieve this, the following tasks will be addressed:

1. Perform a detailed analysis of the raw data, assess its quality, and apply filtering.
2. Develop algorithms for the automatic detection and isolation of objects within images.
3. Investigate the impact of different quality levels of images and spectra on prediction accuracy.
4. Compare the effectiveness of models using single modalities with multimodal approaches.
5. Conduct a comparative study on the publicly available Scene dataset, adapting insights to SDSS in order to validate our multimodal pipeline under controlled conditions.
6. Quantify the relative performance gain of multimodal fusion over unimodal (image-only and spectrum-only) baselines on a structurally similar external dataset to demonstrate the added value of combining modalities.
7. Benchmark and compare training and inference runtimes of all models and modalities on both SDSS and the external dataset, to assess computational scalability and guide practical deployment strategies.

## 1.7 Terminology and Illustrations

### 1.7.1 Spectra and Spectral Analysis

#### 1.7.1.1 Definition of a Spectrum

A spectrum in astronomy represents the dependence of an object's emitted intensity on wavelength. Specialized spectrographs attached to telescopes record these spectra [13].

#### 1.7.1.2 Why Spectral Analysis Is Needed

- **Chemical Composition:** Spectral lines from elements such as hydrogen, oxygen, nitrogen, and iron appear at characteristic wavelengths, and their relative intensities allow us to derive abundances and metallicity in the interstellar medium. For example, the ratio of [O III] to H lines is a common metallicity diagnostic [14]. These abundance measurements are crucial for understanding galactic chemical evolution and enrichment histories [13].
- **Velocity Measurements:** The Doppler shift of spectral lines provides direct measurements of radial velocities, enabling construction of rotation curves and estimates of dynamical mass in galaxies. Line broadening and asymmetries also reveal kinematic components such as outflows, inflows, and turbulent motions [14]. Such velocity diagnostics are essential for probing galaxy dynamics and dark matter distributions.
- **Physical Conditions:** The relative strengths and widths of emission and absorption features encode the temperature, density, and ionization state of the gas. Line ratio diagnostics—such as the [S II] doublet for electron density and the Balmer decrement for dust extinction—help characterize the physical environment within H II regions and around active nuclei [13]. Understanding these conditions informs models of star-formation efficiency and feedback processes.

All of these diagnostics are discussed in Chapter 1 ("Why Record Spectra of Astronomical Objects?") of Tennyson's second edition [13, p. 1–6].

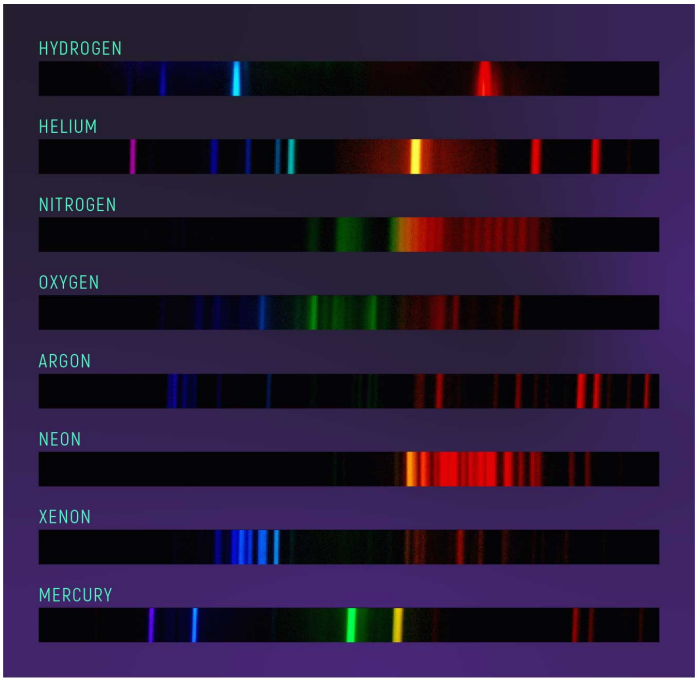



**Figure 1.2** Example of atomic spectral lines for different elements.[15]

### 1.7.2 The SDSS *u, g, r, i, z* Filters

SDSS uses five broadband filters—$u$, $g$, $r$, $i$, and $z$—with effective wavelengths of $u = 354\,\text{nm}$, $g = 477\,\text{nm}$, $r = 623\,\text{nm}$, $i = 762\,\text{nm}$, and $z = 913\,\text{nm}$. Their full-width at half-maximum (FWHM) bandwidths are approximately $\Delta u \approx 56\,\text{nm}$, $\Delta g \approx 138\,\text{nm}$, $\Delta r \approx 138\,\text{nm}$, $\Delta i \approx 152\,\text{nm}$, and $\Delta z \approx 95\,\text{nm}$ [5].

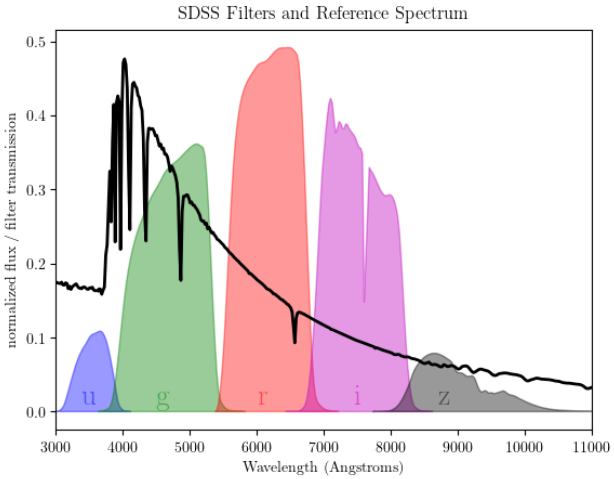



**Figure 1.3** Transmission curves of the SDSS *u, g, r, i, z* filters.

### 1.7.3 Star Formation Rate (SFR)

#### 1.7.3.1 What Is SFR

SFR quantifies the rate of star formation in solar masses per year ($M_\odot\,\mathrm{yr}^{-1}$) [8].

#### 1.7.3.2 How SFR Is Determined

Emission line luminosity, especially H$\alpha$, is used:

$$\mathrm{SFR}(M_\odot\,\mathrm{yr}^{-1}) \approx 7.9 \times 10^{-42}\, L(\mathrm{H}\alpha)\,(\mathrm{erg\,s}^{-1}).$$

[8]

#### 1.7.3.3 Why SFR Is a Key Galactic Parameter

The star formation rate (SFR) underpins multiple aspects of galaxy evolution:

- **Stellar Mass Assembly.** The SFR directly measures the conversion rate of cold gas into stars, driving the build-up of stellar mass and shaping the galaxy stellar mass function over cosmic time [16].
- **Chemical Enrichment.** High SFRs produce core-collapse supernovae and AGB-star mass loss that return heavy elements (e.g., O, Fe) to the interstellar medium, establishing metallicity gradients and enriching subsequent generations of stars [17].
- **Feedback and ISM Regulation.** Radiation pressure, stellar winds, and supernova explosions from young massive stars inject energy and momentum into the ISM, driving turbulence, regulating star formation efficiency, and launching galactic-scale outflows [18].
- **Star Formation Laws.** Empirical relations such as the Kennicutt–Schmidt law relate gas surface density to SFR surface density, providing fundamental insight into the physical processes controlling star formation on galactic and sub-galactic scales [19].
- **Cosmic Star Formation History.** The evolution of the global SFR density with redshift traces galaxy growth, cosmic chemical evolution, and black hole accretion, marking key epochs such as the peak of star formation around $z \sim 2$ and the decline toward the present day [20].

# Chapter 2

# Data Exploration

## 2.1 Dataset Overview and Initial Filtering

We source our sample from the SDSS Data Release 7 star formation rate (SFR) catalog, which initially contains 4 851 200 objects. To ensure that every galaxy has both imaging and spectroscopic data, we retain only those entries with available multi-band cutouts and 1D spectra, reducing the sample to 151 190 records. Next, we remove entries where the logarithmic SFR indicator `AVG` is undefined (NaN), leaving 34 613 objects. Finally, we exclude the placeholder value `AVG` $= -99$, resulting in 30 752 records. Of these, 16 841 have `FLAG`=0 (high-quality SFR estimates) and 13 911 have `FLAG`$\neq$0 [10, 21]. Table 2.1 summarizes these counts.

**Table 2.1** Record counts at successive filtering stages.

| Filtering step | # of Objects |
|---|---|
| Initial SDSS SFR catalog | 4 851 200 |
| With image & spectrum available | 151 190 |
| Removing NaN in `AVG` | 34 613 |
| Excluding `AVG` $= -99$ | 30 752 |
| (`FLAG`=0) | 16 841 |
| (`FLAG`$\neq$0) | 13 911 |

Table 2.1 shows how aggressive filtering reduces the sample to the most reliable SFR measurements for our regression tasks.

Because we leverage the HiSS-Cube framework—a scalable pipeline for hierarchical semi-sparse cubes that preserves measurement uncertainties and precomputes cutouts—each galaxy in our high-quality subset is accompanied by five image quality levels and five spectral resolutions [2]. Moreover, each of these variants carries the same `AVG` SFR label, simplifying our supervised

learning setup.

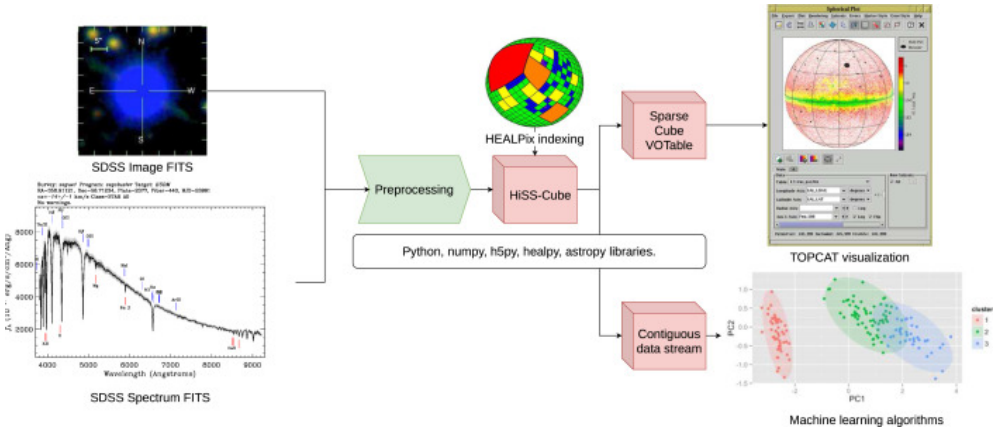



**Figure 2.1** HiSS-Cube data flow pipeline. Starting from SDSS FITS images and spectra (left), uncertainties are extracted and multi-resolution maps generated (Preprocessing). Data are stored in a sparse HDF5 cube indexed by HEALPix for efficient spatial queries. Outputs can be exported as VOTables/FITS for Virtual Observatory tools or as contiguous NumPy arrays for machine learning. The pipeline leverages standard Python libraries (h5py, healpy, astropy) for easy extension.

As illustrated in Fig. 2.1, the HiSS-Cube framework [2] automates the ingestion of SDSS photometric and spectroscopic FITS files, applies preprocessing to generate multi-resolution image cutouts and uniform spectral samplings, and stores everything in a hierarchical sparse cube indexed by HEALPix. This design enables both VO-compatible exports and direct NumPy access for downstream machine learning workflows.

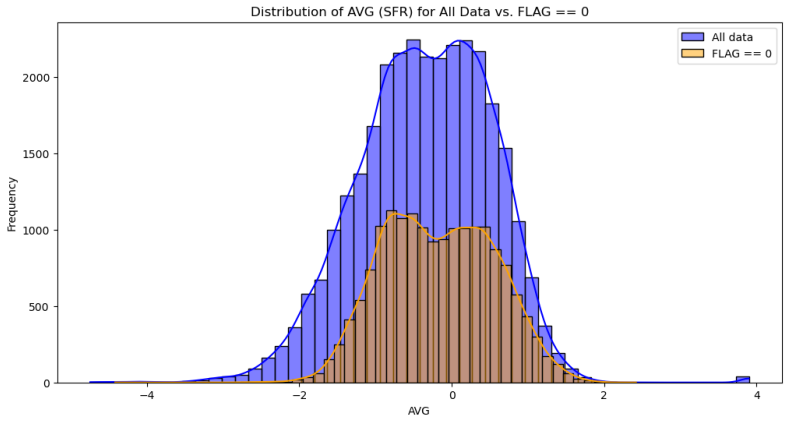



**Figure 2.2** Distribution of `AVG` ($\log_{10}$ SFR) in the filtered sample.

Figure 2.2 reveals a roughly log-normal distribution of SFR values, with most galaxies clustered around $\log_{10}(\mathrm{SFR}) \sim -1.5$ to 0.

## 2.2 Image and Spectrum Data Availability

Thanks to the HiSS-Cube pipeline [2], each high-quality galaxy (`FLAG`=0) is preprocessed into a multi-resolution "cube" that preserves uncertainties. For our regression experiments, we retrieve five image resolutions and five spectral samplings per object (see Fig. **??**).

- **Image cutouts.** Five spatial resolutions with shape $(N, 5, H, W)$, where $H = W \in \{64, 32, 16, 8, 4\}$ pixels. These correspond to successive downsamplings of the original 64×64 cutout, allowing us to study the impact of morphological detail on SFR prediction.

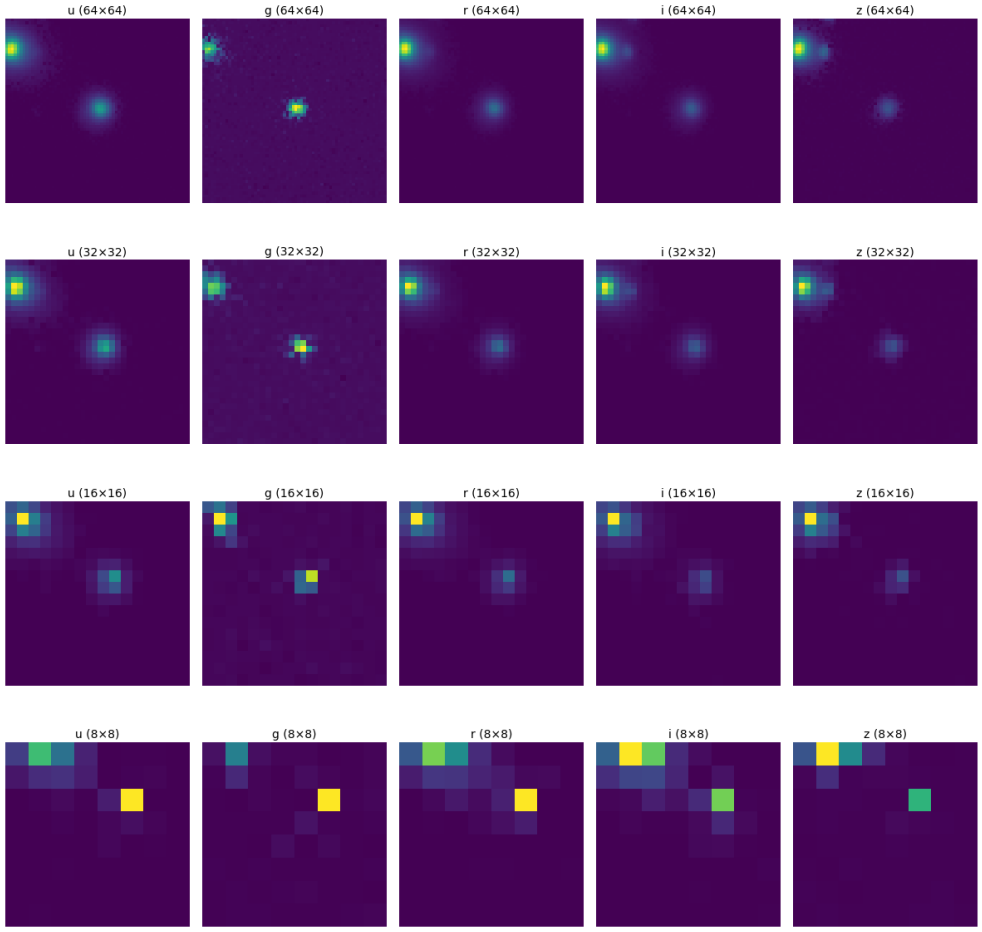



**Figure 2.3** HiSS-Cube image outputs for a single galaxy at five resolution levels (64×64 to 4×4 pixels).

- **Spectral vectors.** Five one-dimensional samplings with length $L \in \{4620, 2310, 1155, 577, 289\}$ bins, obtained by uniform downsampling of the native SDSS spectrum.

Lower-resolution spectra effectively smooth high-frequency noise, serving as a built-in denoiser.

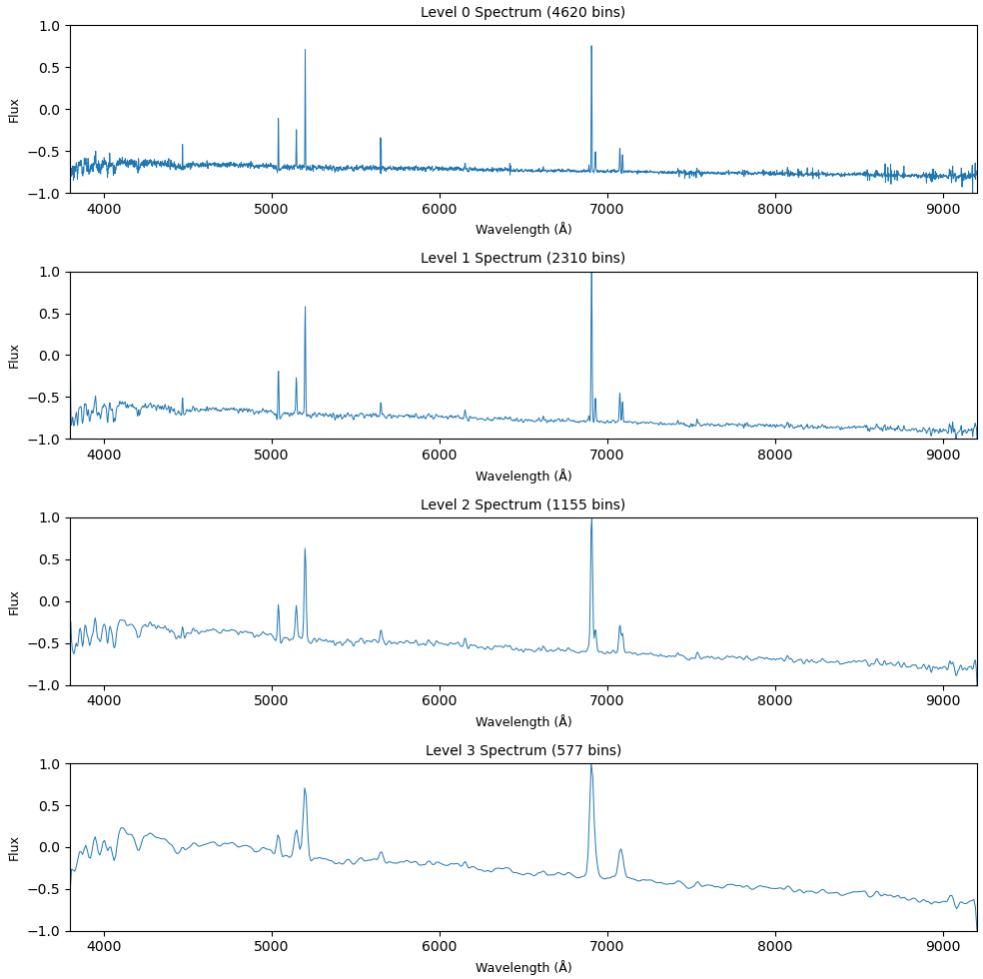



**Figure 2.4** HiSS-Cube spectral outputs for the same galaxy at five sampling levels (4620 to 289 bins).

By having these five distinct quality levels for both images and spectra, we can systematically evaluate how resolution and smoothing affect model performance and computational cost.

## 2.3 SFR Estimation Quality: FLAG Keyword

According to the SDSS documentation:

> "The FLAG keyword indicates the status of the SFR estimation. If FLAG=0 then all is well and for statistical studies in particular, it

is recommendable to focus on these objects as in all other cases the detailed method to estimate SFR or SFR/M* will be (slightly) different and can introduce subtle biases." [10]

We proceed exclusively with the `FLAG`=0 subset (16 841 galaxies).

## 2.4 Analysis of NaN Block Lengths and Positions

### 2.4.1 NaN Percentage by Object

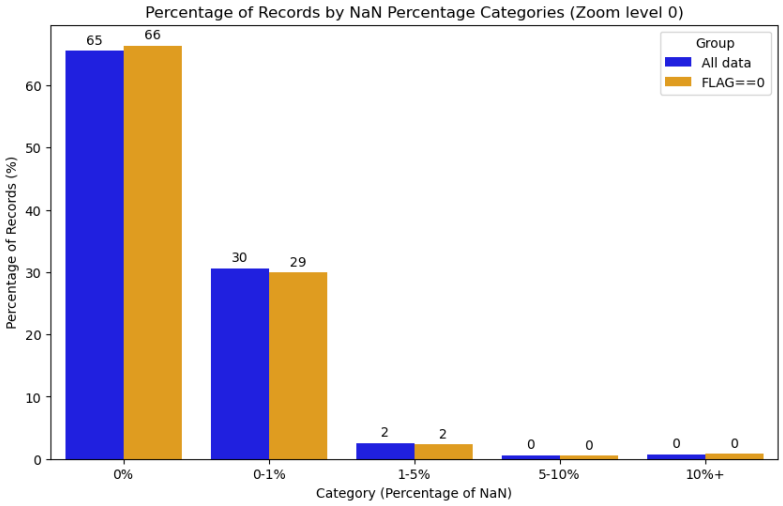



**Figure 2.5** Percentage of records by NaN percentage categories at Zoom level 0, comparing all data vs. `FLAG`=0 subset.

Figure 2.5 shows that over 65% of spectra contain no NaNs, and only about 2% have 1–5% missing values, indicating that most high-quality galaxies have nearly complete spectra.

### 2.4.2 NaN Block Statistics

Before examining spatial patterns, we quantify runs of consecutive NaNs in each spectrum. Table 2.2 reports the total number of NaN blocks, their mean lengths, and maximum lengths at each zoom level.

**Table 2.2** NaN block statistics for `FLAG`=0 at each zoom level.

| Zoom level | # NaN blocks | Mean length | Max length |
|---|---|---|---|
| 0 | 12 207 | 34.69 | 4 620 |
| 1 | 12 045 | 18.11 | 2 310 |
| 2 | 11 954 | 9.68 | 1 155 |
| 3 | 11 875 | 5.46 | 577 |

This table indicates that while the total number of NaN segments is similar across resolutions, the average and maximum block lengths decrease at lower spectral sampling due to downsampling "compressing" gaps.

### 2.4.3 Distribution of NaN Run Lengths

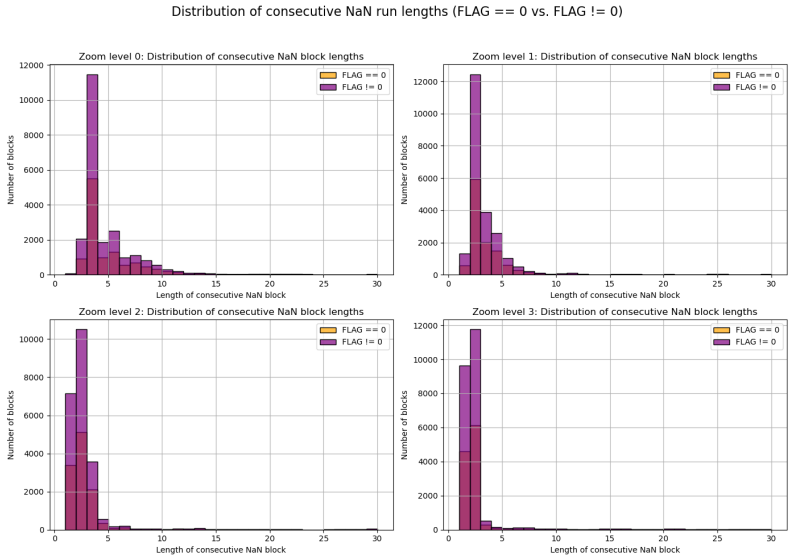



**Figure 2.6** Distribution of consecutive NaN run lengths at each resolution for `FLAG`=0.

In Fig. 2.6, most NaN runs are very short (1–3 bins), with only a few extending beyond 10 bins. This suggests that missing data are typically localized "spikes" rather than large spectral gaps.

### 2.4.4 NaN Occurrence Along Wavelength

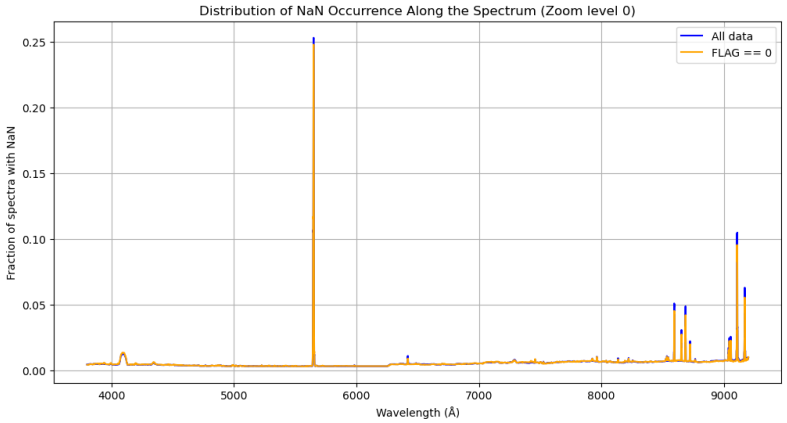
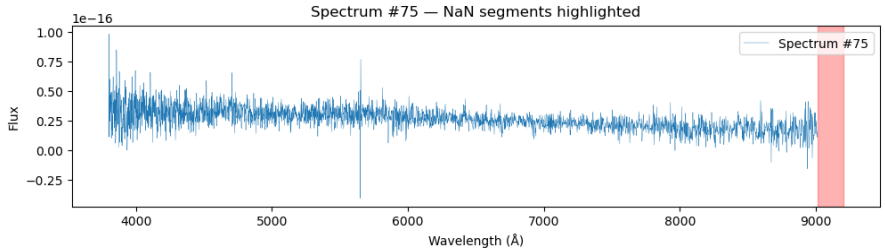


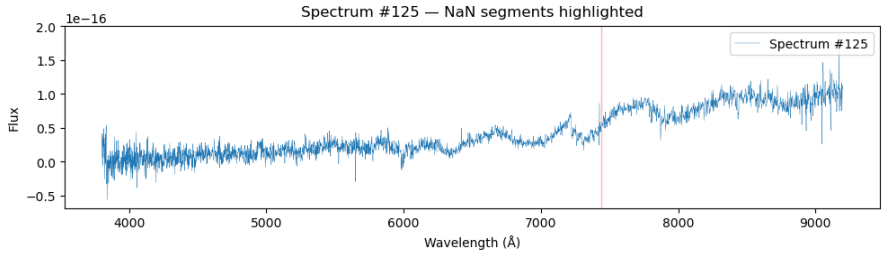

**Figure 2.7** Typical wavelength regions where NaN gaps commonly occur (Zoom level 0).

Figure 2.7 shows peaks in NaN frequency around 5500 Å and near the red end (9000 Å), corresponding to spectrograph join regions and low-sensitivity wavelengths.

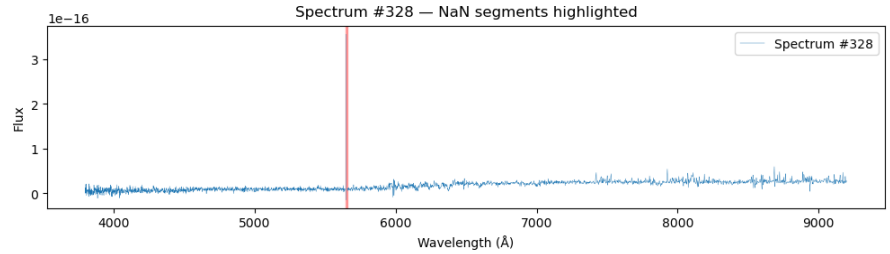
Each point along the wavelength axis represents the fraction of spectra in which that specific bin is flagged as NaN; notably, there is no wavelength where 0% of spectra are missing data, indicating that every channel is affected by occasional dropouts or quality flags. The sharp spike at $\sim 5500\,\text{Å}$ coincides with the dichroic split between the blue and red arms of the SDSS spectrograph, where stitching mismatches and calibration uncertainties often lead to flagged pixels [22]. The elevated NaN occurrence near $\sim 9000\,\text{Å}$ arises from the declining quantum efficiency of the red CCDs and strong telluric emission lines (e.g. atmospheric OH), which reduce the signal-to-noise ratio and trigger data quality filters [23].




**(a)** Spectrum #75. The red shading marks the NaN gap near 9000 Å, where the CCD sensitivity drops.




**(b)** Spectrum #125. The red shading highlights a NaN spike around 7300 Å, likely due to sky–line subtraction residuals.




**(c)** Spectrum #328. The red shading indicates a NaN region near 5600 Å, coincident with the spectrograph arm join.

**Figure 2.8** Examples of SDSS spectra containing NaN segments. In each panel, the red overlay marks the wavelength region flagged as NaN.

## 2.5 Detection and Removal of Multi-Object Cutouts

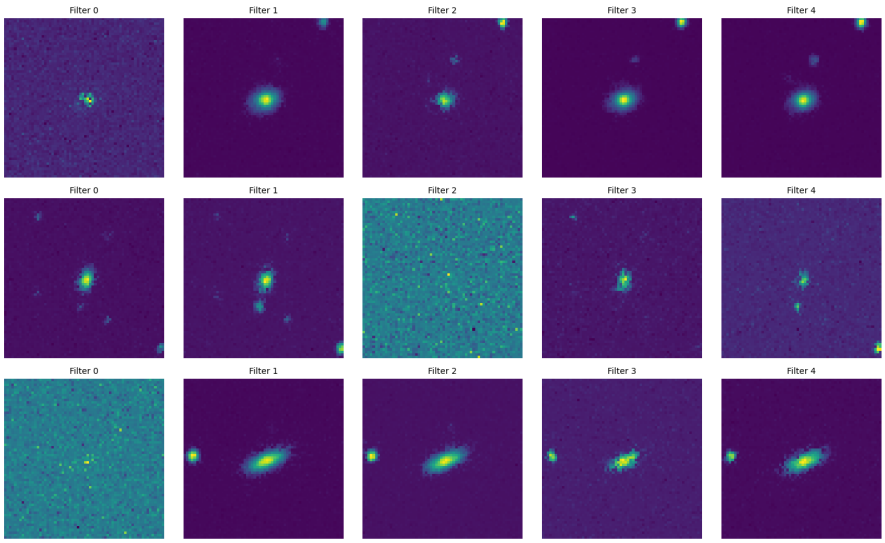



**Figure 2.9** Example of a cutout containing multiple detected sources, excluded from the final sample [21] .

In order to detect and remove cutouts containing multiple objects, we implement a simple image-processing pipeline inspired by standard thresholding and connected-component labeling techniques. First, pixel values are normalized to the [0,1] range. We then binarize the central filter image (usually the $r$-band) at a fixed global threshold of 0.9—this value was chosen heuristically to separate background sky from source signal, following best practices in image thresholding [11]. Next, we apply the connected-component labeling algorithm (‘ndimage.label‘) to the binary image to count discrete regions. If more than one connected region is found, the index is flagged as a “multi-object” cutout. Finally, a small subset of these multi-object indices is visualized to confirm the detection. Our implementation is provided in Listing [21] and closely follows the methodology of Sezgin and Sankur’s survey on thresholding techniques [11] as well as the standard workflow described in Gonzalez and Woods’s digital image processing text [12].

## 2.6 Summary of Final Dataset

The cleaned dataset for supervised regression consists of:

- Multi-band image cutouts at four resolutions
- One-dimensional spectra at four samplings
- Robust SFR labels (`AVG`, `FLAG`=0)

- Total of 11,179 galaxies

### 2.6.1 Exploratory Embedding Analysis with t-SNE, UMAP, and PCA

To gain intuition about the structure of our image and spectral datasets in relation to the target variable `AVG`, we applied three popular dimensionality-reduction methods:

- **t-SNE** [24]
- **UMAP** [25]
- **PCA** [26]

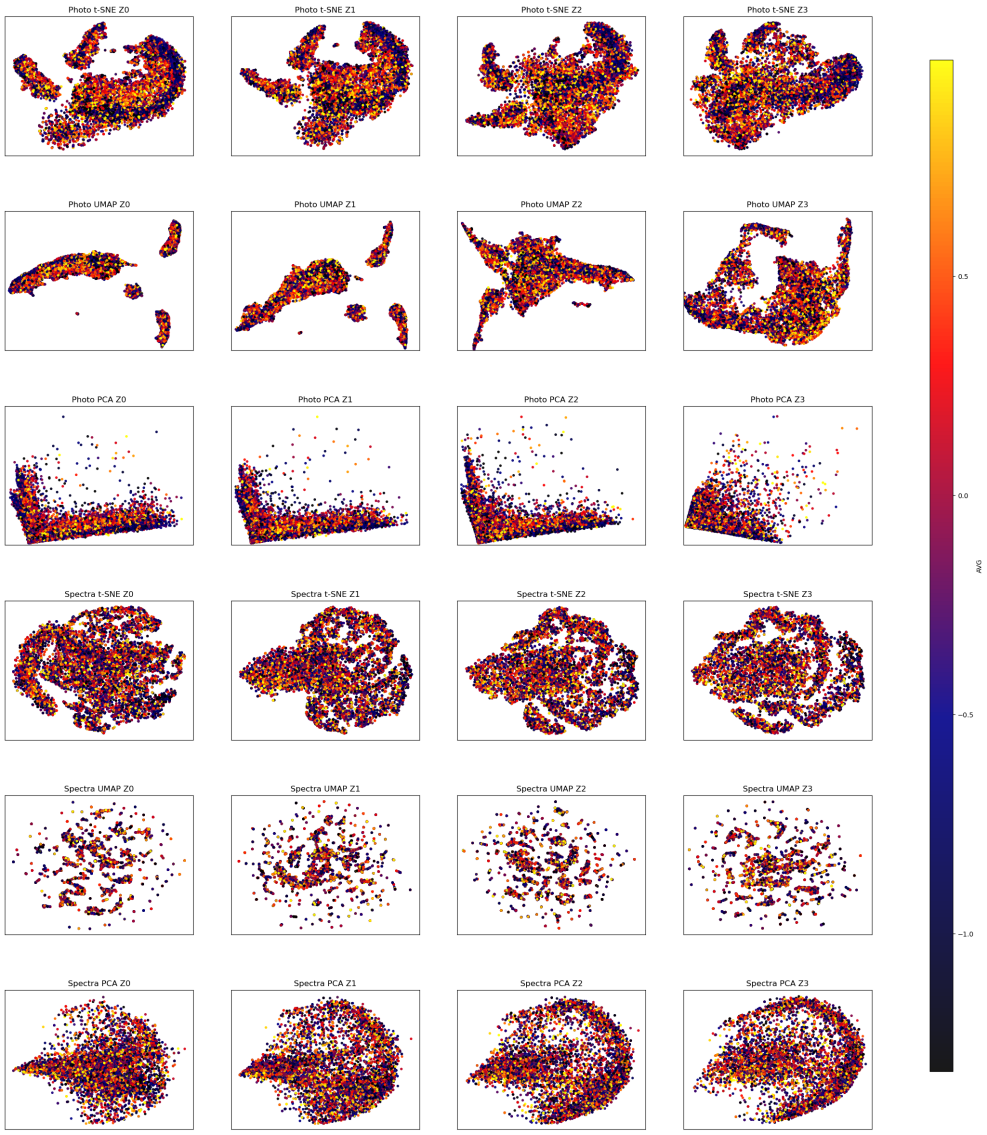



**Figure 2.10** Embeddings of photo and spectral data at four zoom levels (Z0–Z3) using t-SNE, UMAP, and PCA, colored by `AVG` [27].

The Pearson correlations between embedding axes and `AVG` are:

| Method / Modality | $\rho_x$ | $\rho_y$ | Method / Modality | $\rho_x$ | $\rho_y$ |
|---|---|---|---|---|---|
| t-SNE Photo Z0 | -0.03 | -0.04 | t-SNE Spectra Z0 | -0.10 | +0.06 |
| t-SNE Photo Z1 | -0.03 | -0.06 | t-SNE Spectra Z1 | -0.15 | +0.02 |
| t-SNE Photo Z2 | 0.00 | -0.09 | t-SNE Spectra Z2 | -0.16 | +0.00 |
| t-SNE Photo Z3 | -0.06 | -0.03 | t-SNE Spectra Z3 | -0.15 | -0.02 |
| UMAP Photo Z0 | +0.03 | +0.00 | UMAP Spectra Z0 | +0.02 | -0.02 |
| UMAP Photo Z1 | +0.03 | +0.03 | UMAP Spectra Z1 | -0.03 | +0.00 |
| UMAP Photo Z2 | -0.05 | -0.01 | UMAP Spectra Z2 | -0.02 | -0.01 |
| UMAP Photo Z3 | +0.08 | -0.02 | UMAP Spectra Z3 | -0.02 | +0.02 |
| PCA Photo Z0 | -0.03 | +0.01 | PCA Spectra Z0 | -0.11 | +0.05 |
| PCA Photo Z1 | -0.05 | -0.01 | PCA Spectra Z1 | -0.14 | +0.03 |
| PCA Photo Z2 | -0.07 | -0.04 | PCA Spectra Z2 | -0.14 | +0.01 |
| PCA Photo Z3 | -0.07 | +0.03 | PCA Spectra Z3 | -0.14 | -0.01 |

The silhouette scores and cluster-average `AVG` (for $k = 3$) are:

| Method / Modality | Silhouette | Cluster averages of `AVG` |
|---|---|---|
| t-SNE Photo Z0 | 0.43 | [-0.174, -0.169, -0.183] |
| t-SNE Photo Z1 | 0.43 | [-0.189, -0.174, -0.161] |
| t-SNE Photo Z2 | 0.42 | [-0.167, -0.121, -0.238] |
| t-SNE Photo Z3 | 0.44 | [-0.240, -0.108, -0.171] |
| t-SNE Spectra Z0 | 0.40 | [-0.091, -0.160, -0.290] |
| t-SNE Spectra Z1 | 0.41 | [-0.268, -0.174, -0.070] |
| t-SNE Spectra Z2 | 0.39 | [-0.221, -0.244, -0.065] |
| t-SNE Spectra Z3 | 0.37 | [-0.119, -0.283, -0.124] |
| UMAP Photo Z0 | 0.49 | [-0.184, -0.156, -0.218] |
| UMAP Photo Z1 | 0.51 | [-0.221, -0.149, -0.158] |
| UMAP Photo Z2 | 0.43 | [-0.131, -0.267, -0.162] |
| UMAP Photo Z3 | 0.42 | [-0.117, -0.232, -0.183] |
| UMAP Spectra Z0 | 0.37 | [-0.162, -0.153, -0.210] |
| UMAP Spectra Z1 | 0.35 | [-0.190, -0.150, -0.182] |
| UMAP Spectra Z2 | 0.38 | [-0.183, -0.164, -0.179] |
| UMAP Spectra Z3 | 0.34 | [-0.176, -0.194, -0.156] |
| PCA Photo Z0 | 0.54 | [-0.150, -0.201, -0.212] |
| PCA Photo Z1 | 0.56 | [-0.137, -0.245, -0.221] |
| PCA Photo Z2 | 0.58 | [-0.120, -0.295, -0.291] |
| PCA Photo Z3 | 0.62 | [-0.134, -0.303, -0.207] |
| PCA Spectra Z0 | 0.39 | [-0.194, -0.246, -0.056] |
| PCA Spectra Z1 | 0.49 | [-0.286, -0.159, -0.079] |
| PCA Spectra Z2 | 0.49 | [-0.276, -0.075, -0.163] |
| PCA Spectra Z3 | 0.46 | [-0.264, -0.168, -0.070] |

The embedding analysis reveals:

- **t-SNE** and **UMAP** uncover local, nonlinear structure but show weak linear correlation with `AVG`, indicating complex manifold relationships [24, 25].
- **PCA** yields stronger linear gradients in the first component—especially for spectra—suggesting that principal components capture a significant fraction of SFR variance in a linear subspace [26].
- Higher-resolution photos produce higher silhouette scores (up to 0.62 for PCA Photo Z3), indicating clearer cluster separation by SFR labels.

In summary, t-SNE and UMAP highlight nonlinear patterns, while PCA emphasizes linear trends. Combining insights from all three methods guides our feature-engineering and model-selection strategies.

# Chapter 3

# Machine Learning Methodology

## 3.1 Comparative Analysis: The Scene Dataset Example

To preliminarily evaluate the benefits of multimodal learning, we conducted experiments on the publicly available *Scene dataset* [28]. This dataset contains two modalities:

- **Images:** Still frames extracted from videos, each depicting one of eight different environmental scenes.
- **Audio features:** Each image is paired with Mel-Frequency Cepstral Coefficients (MFCCs), representing the corresponding sound context.

The classification task consists of two hierarchical objectives:

- **CLASS1:** Binary classification of the scene as **indoors** or **outdoors**.
- **CLASS2:** Fine-grained classification into one of the eight specific scene types: *classroom, city, river, beach, grocery store, football match, restaurant, forest, jungle.*

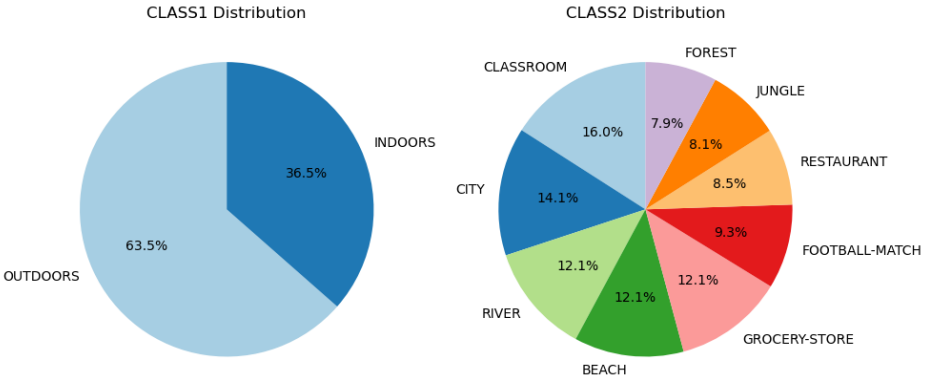



**Figure 3.1** CLASS1 (left) and CLASS2 (right) label distributions for the Scene dataset [29].

During experiments, we observed that prediction accuracy for image-only and multimodal models exceeded 99% for both CLASS1 and CLASS2. Although this suggests strong signal content in the data, it also poses a limitation: the task is too easy to effectively assess the comparative advantage of multimodal learning. In such high-performance regimes, additional modalities do not yield noticeable improvements, making it unsuitable for drawing robust conclusions about fusion strategies.

Therefore, while this dataset helped validate our pipeline, it does not serve as a suitable benchmark for comparing modality contributions. The main focus of this thesis remains on the more challenging SFR prediction task using SDSS data, where both image and spectral inputs contain complementary and non-trivial signals.

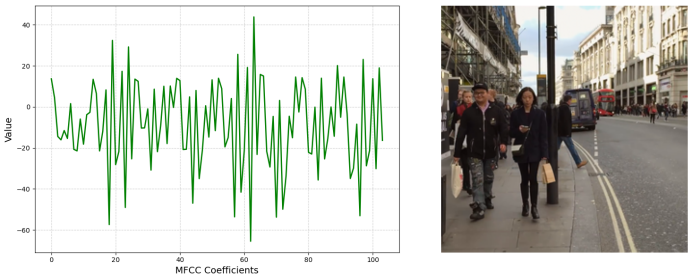



**Figure 3.2** MFCC plot of the audio and street-scene photo taken during the recording [29].

Preliminary machine learning results on this dataset indicate that the multimodal approach significantly improves accuracy:

- **Decision Trees:** Audio-only 0.81/0.66, Combined 0.97/0.92.
- **Neural Networks:** Audio 0.94, Images 0.99, Combined 0.99.

## 3.2 Star–Galaxy–Quasar classification

Unfortunately, attempting a star–galaxy–quasar classification on this dataset proves problematic due to a severe class imbalance. The sample contains roughly ten times more galaxies than quasars, while stars number fewer than 30 instances, making any supervised classifier highly biased toward the majority class. This imbalance stems from the fact that the dataset was originally curated for SFR prediction, not object-type classification.

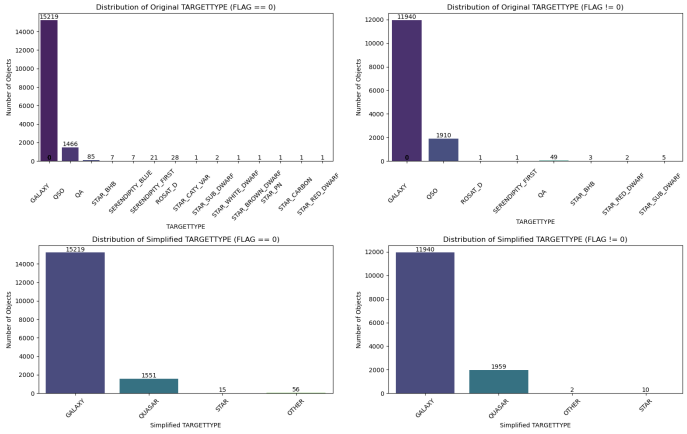



**Figure 3.3** Class distribution for star–galaxy–quasar labels: galaxies outnumber quasars by a factor of 10, and stars comprise fewer than 30 objects [21].

## 3.3 Overview of Learning Algorithms

To predict the logarithmic star-formation rate (`AVG` in $[-4, 4]$) we employ three baseline models:

- **Decision Tree Regression (DT).** A non-parametric tree model that recursively partitions feature space by axis-aligned splits, offering interpretability and a natural baseline [30].
- **Convolutional Neural Network (VGGNet12).** A 12-layer CNN architecture that excels at large-scale image feature extraction [31].
- **Gradient Boosting Machine (LightGBM).** An efficient implementation of gradient-boosted decision trees optimized for speed and memory [32].

## 3.4 Experimental Setup

### 3.4.1 Data Splitting Strategy

We shuffle and split the cleaned sample into training, validation, and test subsets in a 60/20/20 ratio using stratified sampling on `AVG`. We then perform 5-fold cross-validation on the training set to estimate generalization error and tune hyperparameters [33, 34].

### 3.4.2 Preprocessing

- *Images:* pixel values are linearly scaled to $[0, 1]$ by dividing by 255 [35], then flattened for decision-tree/LightGBM models or fed as 2D arrays into VGGNet12 [36].
- *Spectra:* Any object with NaN flux values removed, yielding 11,179 gap-free spectra[37].
- *Early Fusion:* Concatenate image and spectral vectors into one feature vector [38].
- *Late Fusion:* Average photo-only and spec-only model predictions[38].

### 3.4.3 Hyperparameter Tuning

**DT:** grid search over `max_depth` $\in \{1, \ldots, 6\}$ with 5-fold CV, selecting the depth maximizing mean test $R^2$ [30].

**VGGNet12:** sweep over learning rate (`lr`) and fixed dropout=0.5, early stopping patience=30 [39, 40].

**LightGBM:** grid over `learning_rate` and `max_depth`, early stopping round=10 [41].

## 3.5 Evaluation Metrics

We evaluate all models using:

- *Coefficient of Determination ($R^2$).* Variance explained [30].

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}.$$

- *Mean Absolute Error (MAE).* Average absolute deviation [30].

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|.$$

- *Root Mean Square Error (RMSE).* Quadratic penalty on large errors [30].

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}.$$

- *Normalized Median Absolute Deviation (NMAD).* $1.4826 \times \text{median}(|\epsilon - \text{median}(\epsilon)|)$ [42].

$$\text{NMAD} = 1.4826 \times \text{median}\big(|\epsilon_i - \text{median}(\epsilon)|\big), \quad \epsilon_i = y_i - \hat{y}_i.$$

## 3.6 Multimodal Fusion Strategies

### 3.6.1 Early Fusion

Concatenate CNN feature vector (size $N_{\text{img}}$) with spectral vector (size $N_{\text{spec}}$) into one regressor input [43].

### 3.6.2 Late Fusion

Average independent predictions:

$$\hat{y}_{\text{late}} = \tfrac{1}{2}\big(\hat{y}_{\text{photo}} + \hat{y}_{\text{spec}}\big).$$

## 3.7 Decision Tree Regression

We fit DT regressors of depth 1–6 to photo, spectra, and early-fused data, then average photo and spectra for late fusion.

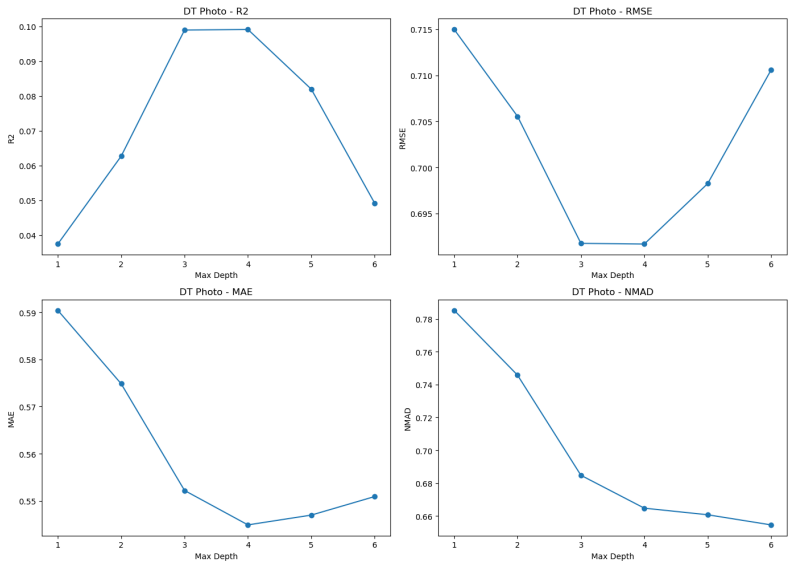



**Figure 3.4** DT on photographs: $R^2$, MAE, RMSE, and NMAD vs. max. tree depth. Best $d = 4$ (all except NMAD).

**Figure 1:** Photo-only DT performance.

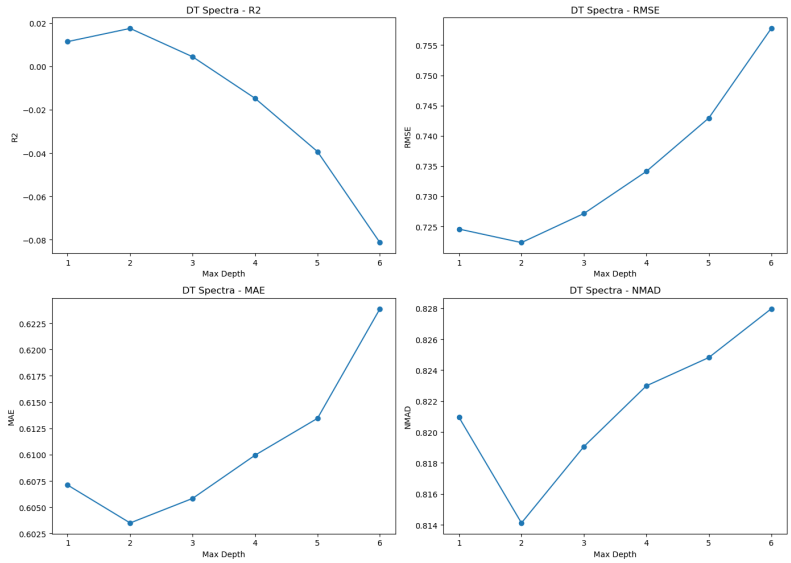



**Figure 3.5** DT on spectra: $R^2$, MAE, RMSE, and NMAD vs. max. tree depth. Best $d = 2$.

**Figure 2:** Spectra-only DT performance.

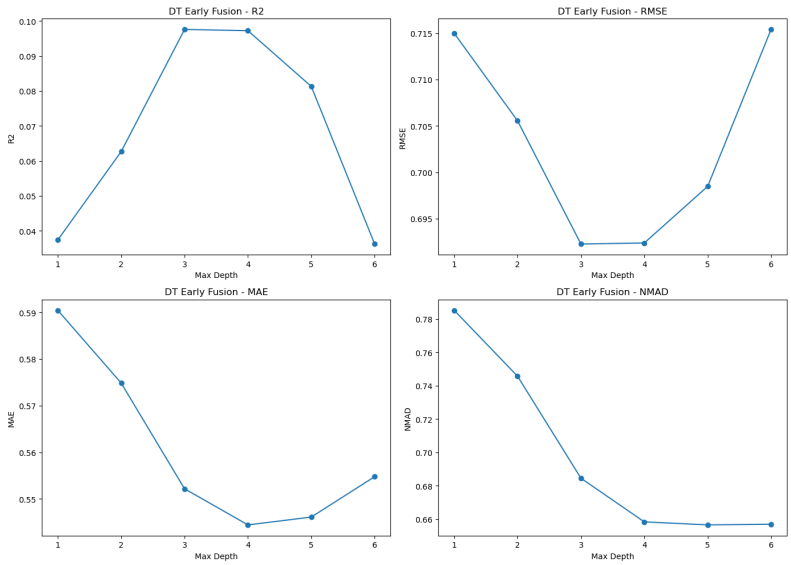



**Figure 3.6** DT early fusion: $R^2$, MAE, RMSE, and NMAD vs. tree depth. Best $d = 3$ by $R^2$.

**Figure 3:** Early fusion DT performance.

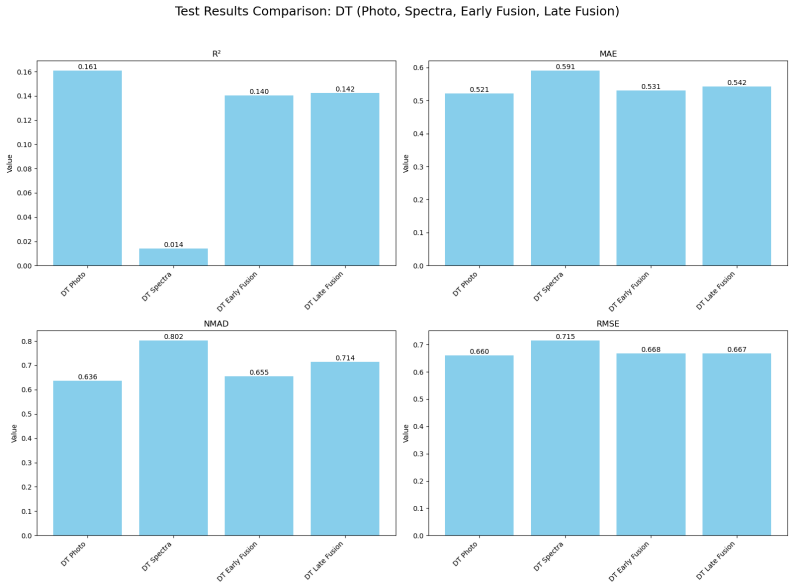



**Figure 3.7** DT: metric comparison across modalities (photo, spectra, early, late).

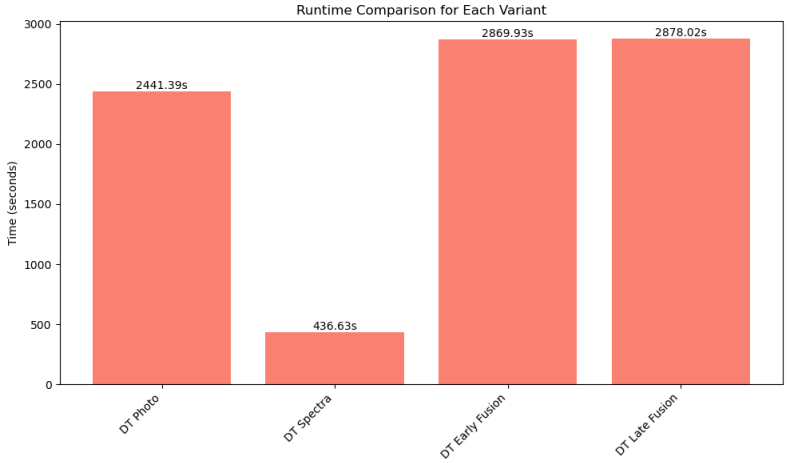



**Figure 3.8** DT: wall-clock runtime across modalities.

## 3.8 Convolutional Neural Network: VGGNet12

The VGGNet12 model stacks $3 \times 3$ convolutions, max-pooling, then three FC layers with dropout, fine-tuned from ImageNet [31].

### 3.8.1 Architecture and Training Protocol

We optimize custom MSE loss,

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2,$$

using Adam, early stopping (patience=30), and focus hyperparameter tuning on learning rate [44, 40, 39].

### 3.8.2 Training Curves: Photographs

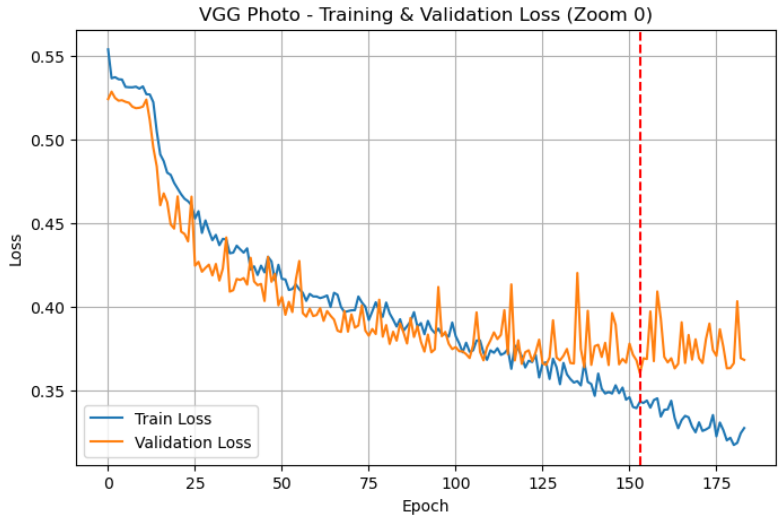



**Figure 3.9** VGGNet12 photo: training (blue) vs. validation (orange) loss per epoch; red dashed line marks lowest val. loss.

**Best params (photo):** { `lr`: `1e-05`, `dropout`: 0.5 }

### 3.8.3 Hyperparameter Sweep: Photographs

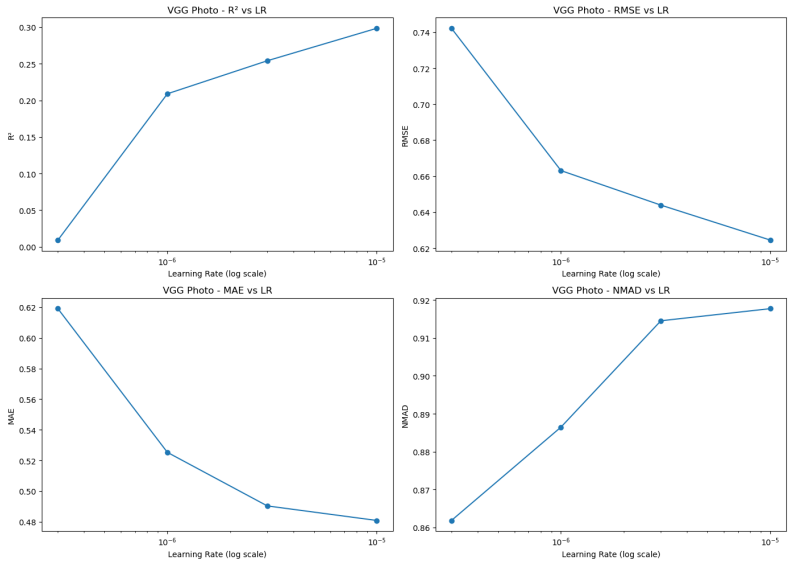



**Figure 3.10** VGGNet12 photo: $R^2$, MAE, RMSE, NMAD vs. learning rate.

**Best params (photo):** { `lr`: `1e-05`, `dropout`: 0.5 }

### 3.8.4 Training Curves: Spectra

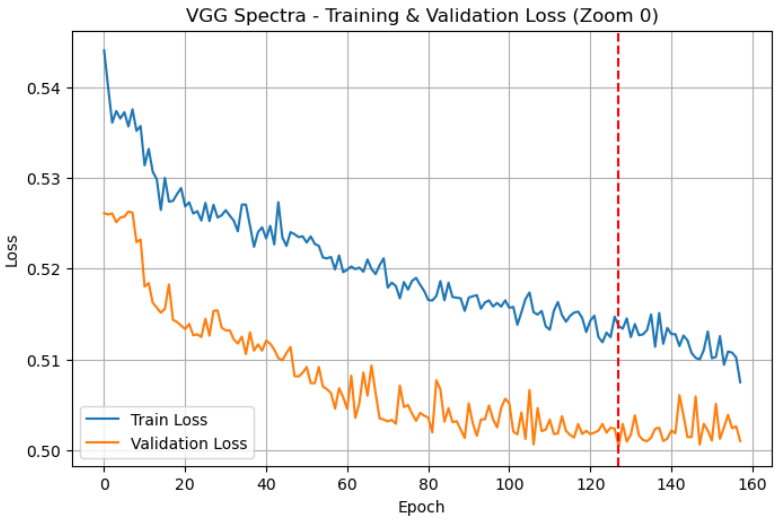



**Figure 3.11** VGGNet12 spectra: training vs. validation loss per epoch; red dashed line = best epoch.

**Best params (spectra):** { `lr`: `3e-06`, `dropout`: 0.5 }

On this graph, it is particularly noticeable that there are epochs where the validation loss dips below the training loss. This behavior is expected in networks using dropout: during training dropout with $p = 0.5$ randomly deactivates neurons, adding noise and raising training loss, whereas no dropout is applied during validation, so the validation loss can occasionally be lower than the training loss [45].

### 3.8.5 Hyperparameter Sweep: Spectra

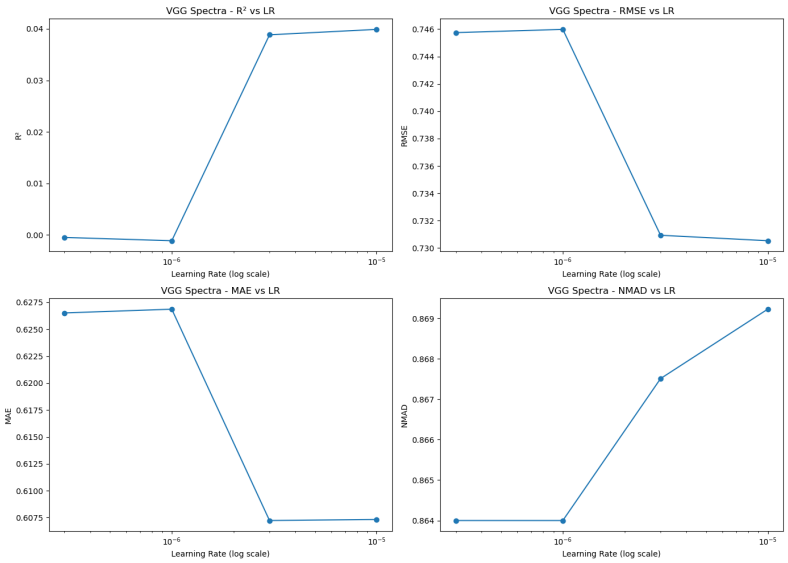



**Figure 3.12** VGGNet12 spectra: $R^2$, MAE, RMSE, NMAD vs. learning rate.

**Best params (spectra):** { `lr`: `3e-06`, `dropout`: 0.5 }

### 3.8.6 Training Curves: Early Fusion

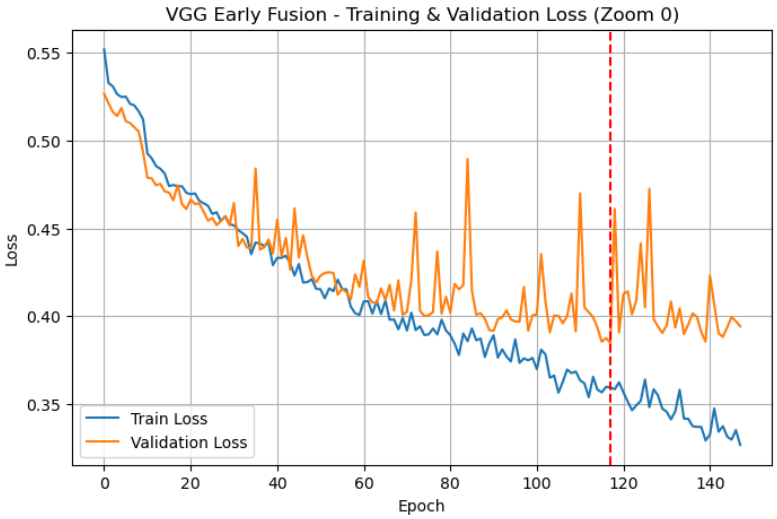



**Figure 3.13** VGGNet12 early fusion: training vs. validation loss; red dashed line = best epoch.

**Best params (early fusion):** { `lr`: `1e-05`, `dropout`: 0.5 }

### 3.8.7 Hyperparameter Sweep: Early Fusion

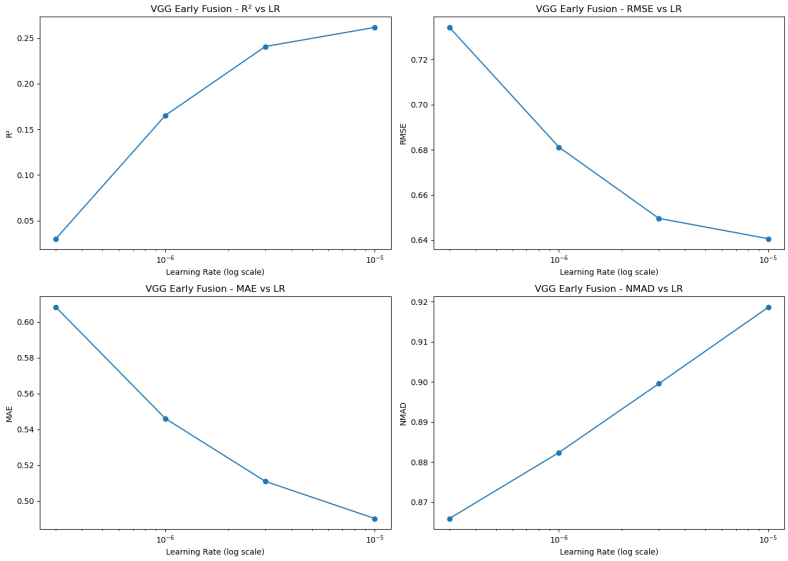



**Figure 3.14** VGGNet12 early fusion: $R^2$, MAE, RMSE, NMAD vs. learning rate.

**Best params (early fusion):** { `lr`: `1e-05`, `dropout`: 0.5 }

### 3.8.8 Overall Metrics and Runtime

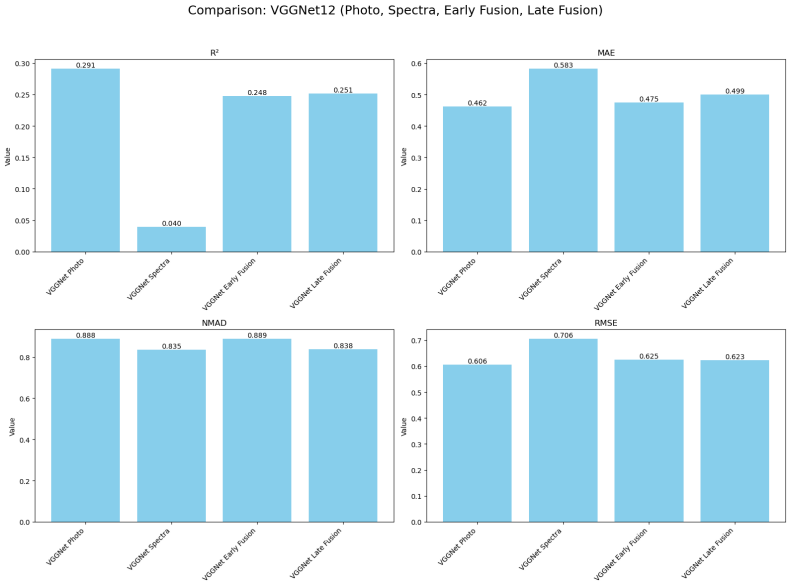



**Figure 3.15** VGGNet12: metric comparison across modalities.

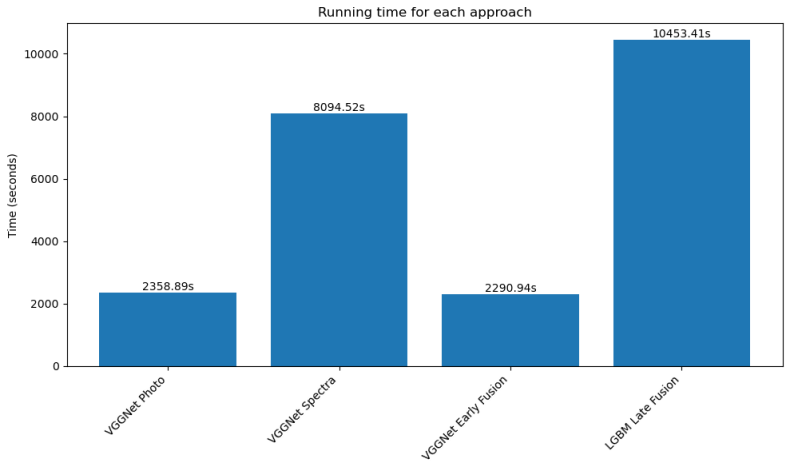



**Figure 3.16** VGGNet12: wall-clock runtime across modalities.

## 3.9 Gradient Boosting Machine: LightGBM

LightGBM grows trees leaf-wise with histogram-based splitting and optimizes RMSE with early stopping (10 rounds) [32, 41].

### 3.9.1 Architecture and Training Protocol

We minimize RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2},$$

and tune `learning_rate` and `max_depth`; early stopping prevents overfitting [46].

### 3.9.2 Training Curves: Photographs

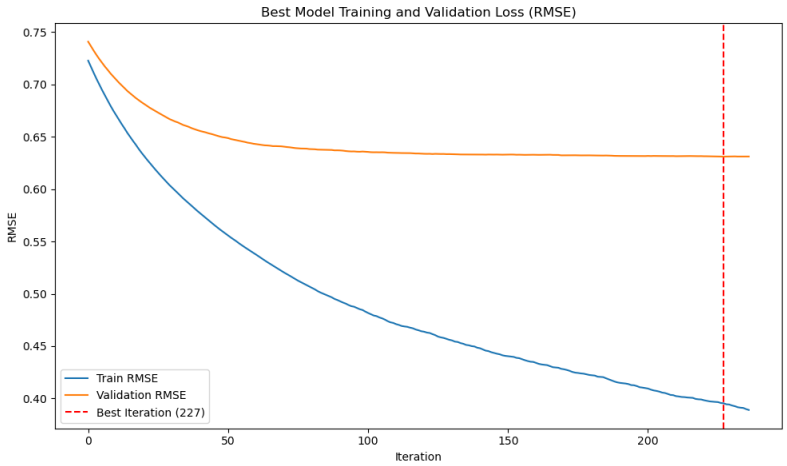



**Figure 3.17** LightGBM photo: training vs. validation RMSE per iteration; red dashed line = best iteration.

**Best params (photo):** { `learning_rate`: 0.1, `max_depth`: 8 }

### 3.9.3 Hyperparameter Sweep: Photographs

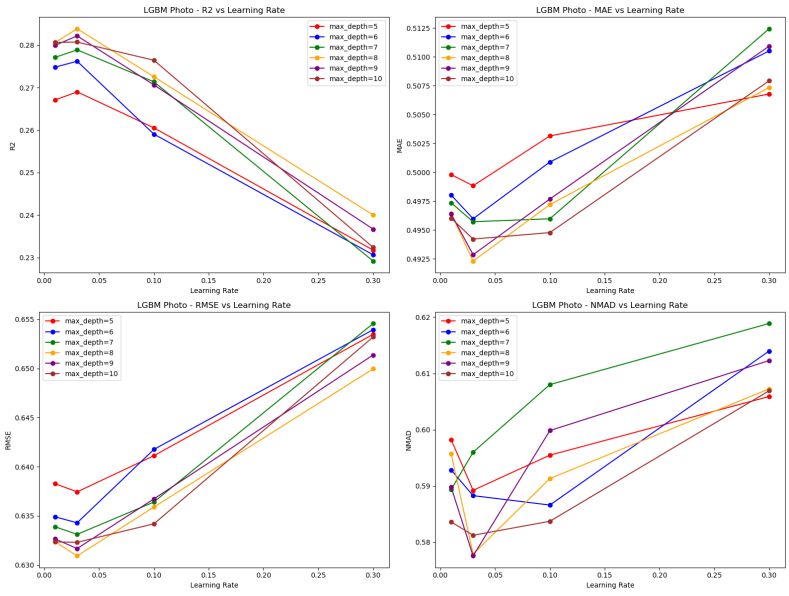



**Figure 3.18** LightGBM photo: $R^2$, MAE, RMSE, NMAD vs. learning rate & max_depth.

**Best params (photo):** { `learning_rate`: 0.1, `max_depth`: 8 }

### 3.9.4 Training Curves: Spectra

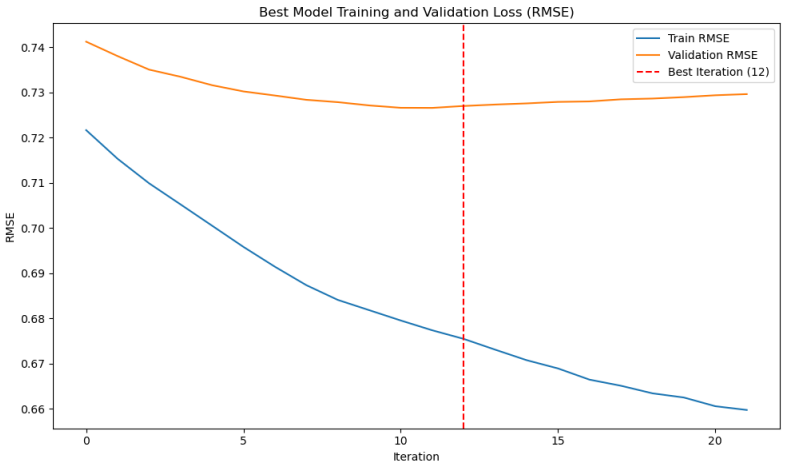



**Figure 3.19** LightGBM spectra: training vs. validation RMSE; red dashed line = best iteration.

**Best params (spectra):** { `learning_rate`: 0.03, `max_depth`: 7 }

### 3.9.5 Hyperparameter Sweep: Spectra

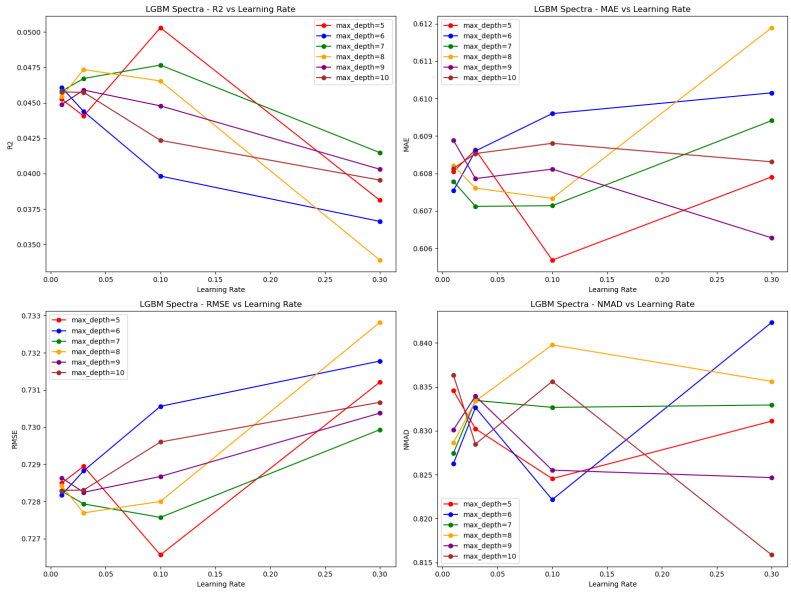



**Figure 3.20** LightGBM spectra: $R^2$, MAE, RMSE, NMAD vs. learning rate & max_depth.

**Best params (spectra):** { `learning_rate`: 0.03, `max_depth`: 7 }

### 3.9.6 Training Curves: Early Fusion

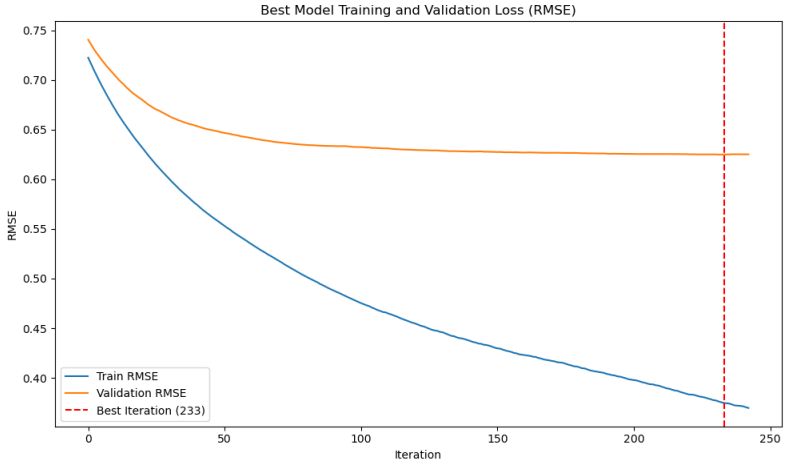



**Figure 3.21** LightGBM early fusion: training vs. validation RMSE; red dashed line = best iteration.

**Best params (early fusion):** { `learning_rate`: 0.1, `max_depth`: 9 }

### 3.9.7 Hyperparameter Sweep: Early Fusion

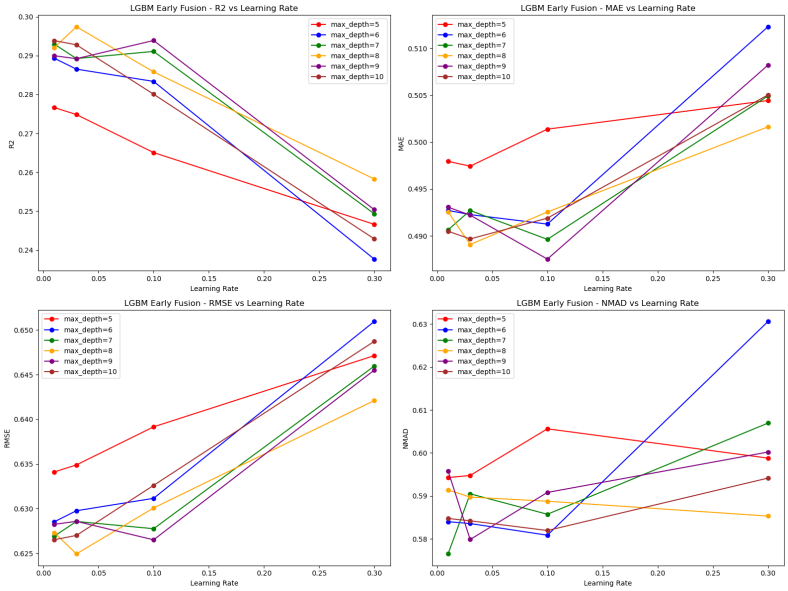



**Figure 3.22** LightGBM early fusion: $R^2$, MAE, RMSE, NMAD vs. learning rate & max_depth.

**Best params (early fusion):** { `learning_rate`: 0.1, `max_depth`: 9 }

### 3.9.8 Overall Metrics and Runtime

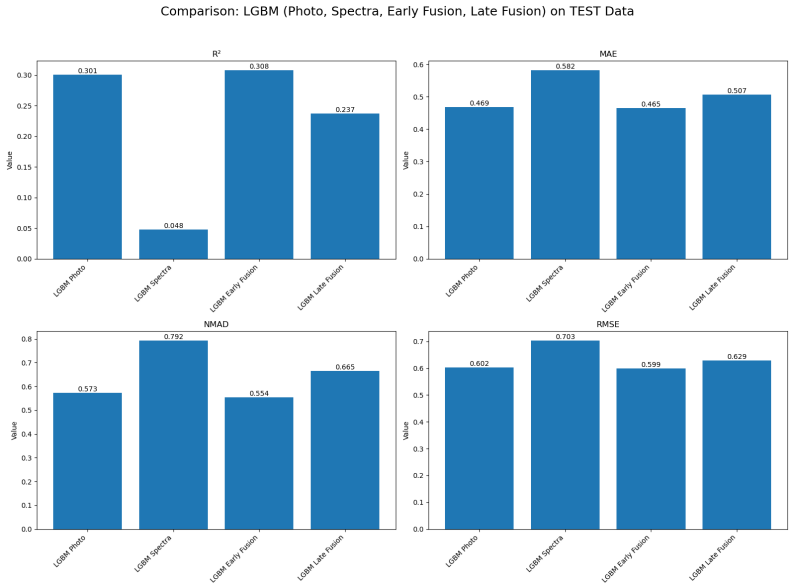



**Figure 3.23** LightGBM: metric comparison across modalities.

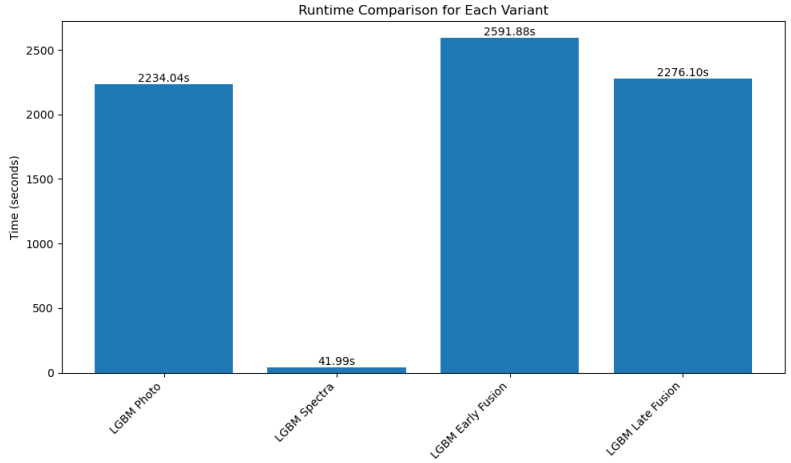



**Figure 3.24** LightGBM: wall-clock runtime across modalities.

## 3.10 Impact of Photo and Spectra Quality on Model Performance

To understand how input quality affects our models, we trained each algorithm separately on all four photo-quality and four spectra-quality variants using Decision Trees, VGGNet12, and LightGBM. Figures 3.25, 3.26 and 3.27 summarize the photo results, and Figures 3.28, 3.29 and 3.30 the spectra results.

For **photographs**, all metrics improve monotonically with image quality: higher resolution yields higher $R^2$ and lower MAE, RMSE, and NMAD, at the cost of longer training time.

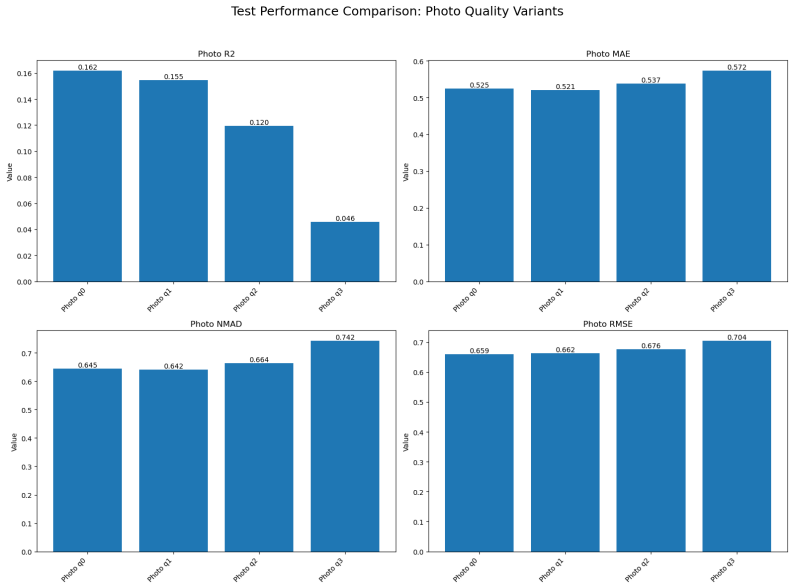



**Figure 3.25** Decision Tree performance vs. photo quality (q0–q3) [47].

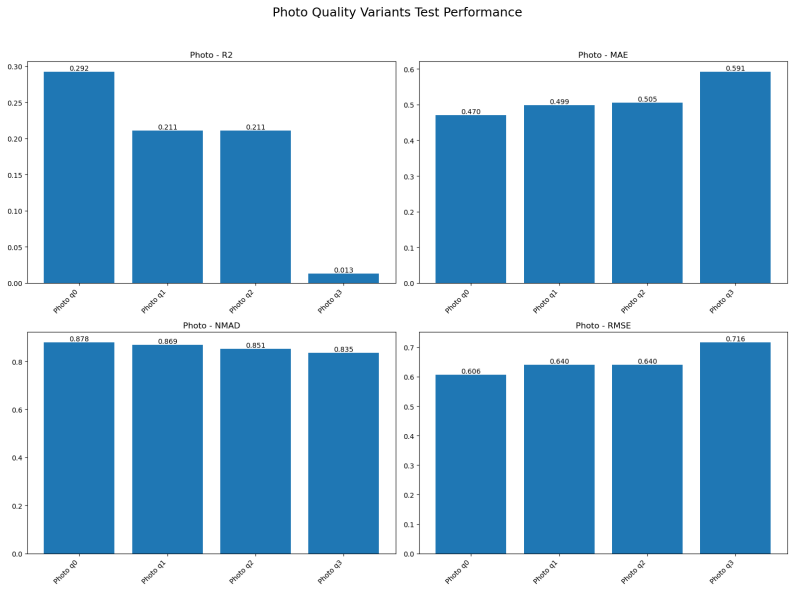



**Figure 3.26** VGGNet12 performance vs. photo quality (q0–q3) [48].

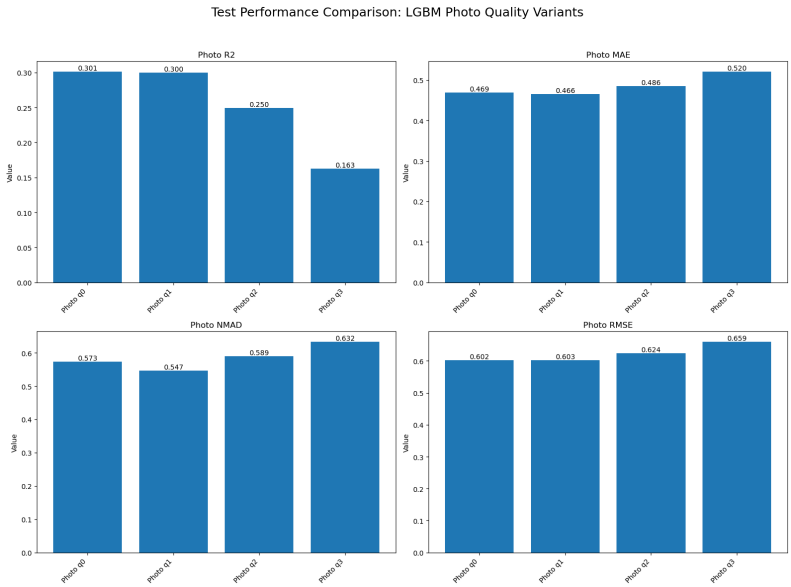



**Figure 3.27** LightGBM performance vs. photo quality (q0–q3) [49].

For **spectra**, the trend is inverted: the lowest-resolution spectra produce the best regression accuracy. We attribute this to the smoothing effect of down-sampling, which attenuates high-frequency noise and acts like a built-in Savitzky–Golay filter, improving generalization [50]. Lower-quality variants also run faster.

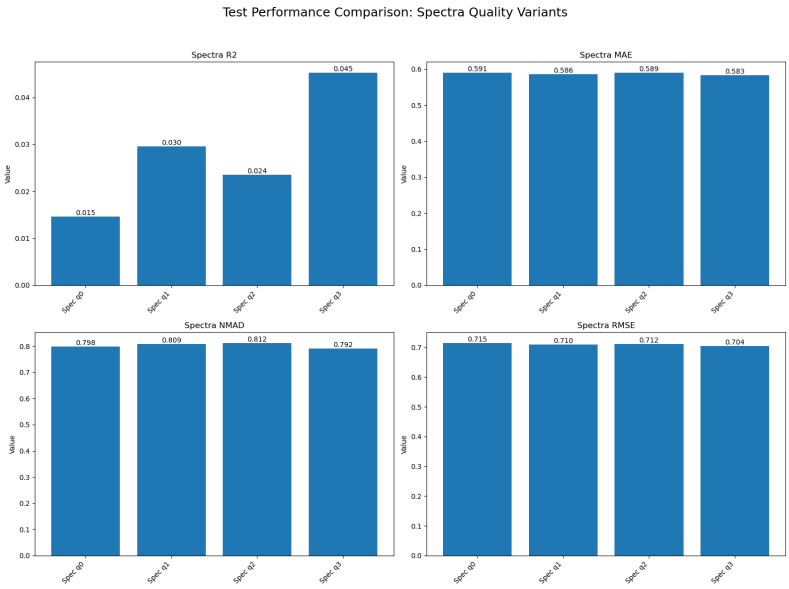



**Figure 3.28** Decision Tree performance vs. spectra quality (q0–q3) [47].

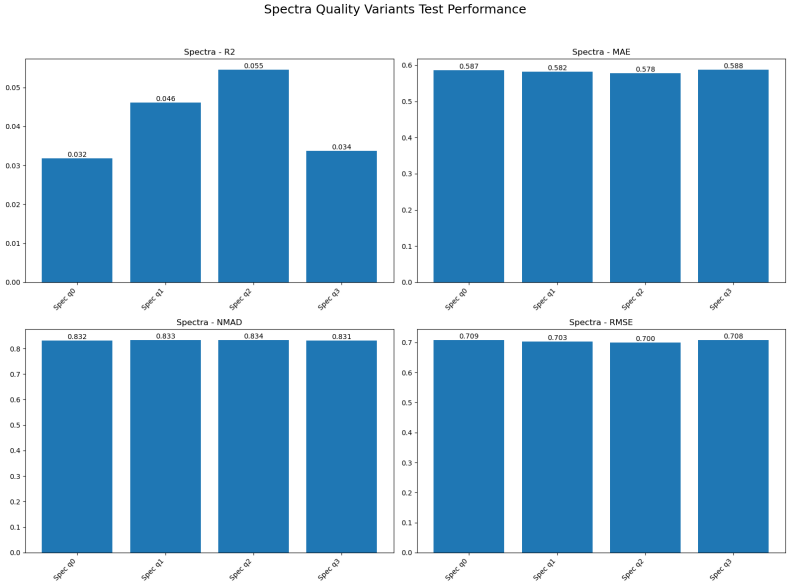



**Figure 3.29** VGGNet12 performance vs. spectra quality (q0–q3) [48].

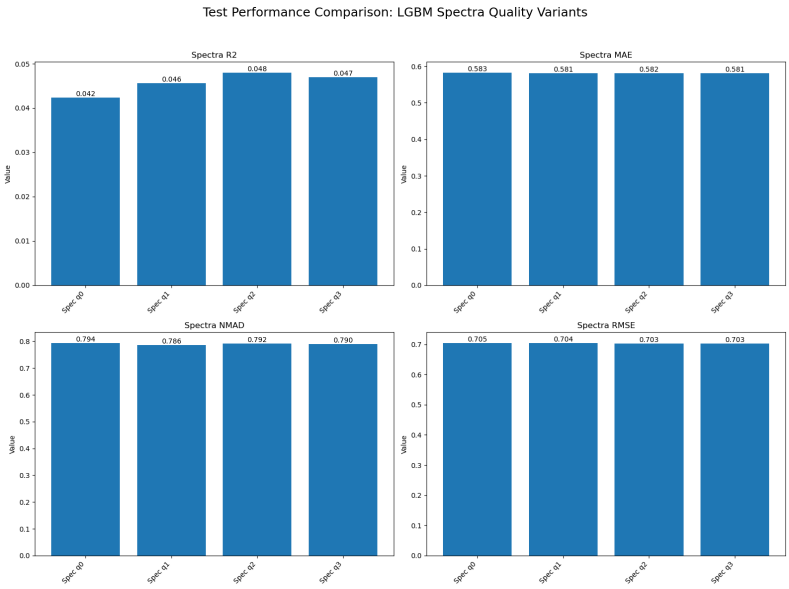



**Figure 3.30** LightGBM performance vs. spectra quality (q0–q3) [49].

Based on these insights, we re-ran our final multimodal experiments using the highest photo quality with the lowest spectra quality for each model. On the test set, the early-fusion $R^2$ changed from

$$(\text{DT} : 0.140,\ \text{VGG} : 0.248,\ \text{LGBM} : 0.308) \quad \rightarrow \quad (\text{DT} : 0.155,\ \text{VGG} : 0.262,\ \text{LGBM} : 0.308),$$

and late-fusion from

$$(\text{DT} : 0.142,\ \text{VGG} : 0.251,\ \text{LGBM} : 0.237) \quad \rightarrow \quad (\text{DT} : 0.160,\ \text{VGG} : 0.262,\ \text{LGBM} : 0.237).$$

These small but consistent gains confirm that moderate smoothing of spectral inputs can enhance multimodal performance.

## 3.11 Summary and Outlook

Among all models and fusion strategies evaluated, the early-fusion LightGBM model achieved the best overall performance across metrics (highest $R^2$, lowest MAE and RMSE), which is consistent with the literature showing that combining complementary modalities at the feature level often yields superior predictive power [51]. Additionally, VGGNet12 applied to photometric images alone performed remarkably well, underscoring the strength of deep CNN feature extractors for morphological information in galaxy images [52].

These results demonstrate that multimodal approaches—particularly early fusion with efficient tree-based learners—can capture both spectral and visual cues essential for accurate SFR prediction. However, the complexity and diversity of astrophysical data suggest that further research is needed: exploring

larger ensembles of models, advanced fusion techniques (e.g., attention-based or late-stage meta-learners), and integration of additional modalities (e.g., environmental or kinematic data) could drive even better performance.

Overall, this work establishes a solid methodological foundation for predicting galaxy star formation rates using multimodal ML, and points the way toward deeper investigations that leverage state-of-the-art models and richer datasets in future studies.

# Bibliography


1. *SDSS Data Release 7* [online]. [N.d.]. Available also from: `https://classic.sdss.org/dr7/`. [accessed 2025-04-28].
2. NÁDVORNÍK, Jirí; ŠKODA, P; TVRDÍK, Pavel. HiSS-cube: A scalable framework for hierarchical semi-sparse cubes preserving uncertainties. *Astronomy and Computing*. 2021, vol. 36, p. 100463.
3. YORK, Donald G; ADELMAN, Jennifer; ANDERSON JR, John E; ANDERSON, Scott F; ANNIS, James; BAHCALL, Neta A; BAKKEN, JA; BARKHOUSER, Robert; BASTIAN, Steven; BERMAN, Eileen, et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal*. 2000, vol. 120, no. 3, p. 1579.
4. LOPES, Amanda R; TELLES, Eduardo; MELNICK, Jorge. The effects of star formation history in the SFR–M* relation of H ii galaxies. *Monthly Notices of the Royal Astronomical Society*. 2021, vol. 500, no. 3, pp. 3240–3253.
5. FUKUGITA, M; SHIMASAKU, K; ICHIKAWA, T; GUNN, JE, et al. *The Sloan digital sky survey photometric system*. 1996. Tech. rep. SCAN-9601313.
6. ABAZAJIAN, K.; ADELMAN-MCCARTHY, J. K.; AGÜEROS, M. A.; *ET AL.* The Seventh Data Release of the Sloan Digital Sky Survey. *Astrophys. J. Suppl. Ser.* 2009, vol. 182, pp. 543–558. Available from DOI: `10.1088/0067-0049/182/2/543`.
7. ALBARETI, Franco D; PRIETO, Carlos Allende; ALMEIDA, Andres; ANDERS, Friedrich; ANDERSON, Scott; ANDREWS, Brett H; ARAGÓN-SALAMANCA, Alfonso; ARGUDO-FERNÁNDEZ, Maria; ARMENGAUD, Eric; AUBOURG, Eric, et al. The 13th data release of the Sloan Digital Sky Survey: First spectroscopic data from the SDSS-IV survey mapping nearby galaxies at Apache Point Observatory. *The Astrophysical Journal Supplement Series*. 2017, vol. 233, no. 2, p. 25.

8. KENNICUTT JR, Robert C. Star formation in galaxies along the Hubble sequence. *Annual Review of Astronomy and Astrophysics*. 1998, vol. 36, no. 1, pp. 189–231.

9. MPA GARCHING. *Raw data* [online]. 2007. Available also from: `https://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/raw_data.html`. [accessed 2025-04-21].

10. MPA GARCHING. SDSS DR7 SFR documentation. *MPA Garching Web Resource*. 2007. Available also from: `https://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/sfrs.html`.

11. SEZGIN, Mehmet; SANKUR, Bu
lent. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging*. 2004, vol. 13, no. 1, pp. 146–168.

12. GONZALEZ, Rafael C. *Digital image processing*. Pearson education india, 2009.

13. TENNYSON, Jonathan. *Astronomical spectroscopy: An introduction to the atomic and molecular physics of astronomical spectroscopy*. World Scientific, 2019.

14. OSTERBROCK, Donald E; FERLAND, Gary J. *Astrophysics Of Gas Nebulae and Active Galactic Nuclei*. University science books, 2006.

15. INSTITUTE, Space Telescope Science. *Spectroscopy 101 – Types of Spectra and Spectroscopy — Webb* [online]. 2022. Available also from: `https://webbtelescope.org/contents/articles/spectroscopy-101--types-of-spectra-and-spectroscopy?page=1&keyword=Stars`. [accessed 2025-04-22].

16. KENNICUTT, Robert C. Jr; EVANS, Neal J. Star Formation in the Milky Way and Nearby Galaxies. *Annu. Rev. Astron. Astrophys.* 2012, vol. 50, pp. 531–608. Available from DOI: `10.1146/annurev-astro-081811-125610`.

17. PRANTZOS, N. Nucleosynthesis in Stars and the Chemical Enrichment of Galaxies. *Annu. Rev. Astron. Astrophys.* 2013, vol. 51.

18. RUPKE, David S. N. A Review of Recent Observations of Galactic Winds Driven by Star Formation. *Galaxies*. 2018, vol. 6, no. 4, p. 114.

19. KENNICUTT, Robert C. The Global Schmidt Law in Star-forming Galaxies. *Astrophys. J.* 1998, vol. 498, pp. 541–552.

20. MADAU, Piero; DICKINSON, Mark. Cosmic Star-Formation History. *Annu. Rev. Astron. Astrophys.* 2014, vol. 52, pp. 415–486. Available from DOI: `10.1146/annurev-astro-081811-125615`.

21. RUSTAMOV, Farukh. *Jupyter Notebook: ≪data$_e$xploring ≫* [online]. 2025. [accessed 2025-04-22].

22. GUNN, J. E.; SIEGMUND, W. A.; MANNERY, E. J.; *ET AL.* The 2.5 m Telescope of the Sloan Digital Sky Survey. *Astron. J.* 2006, vol. 131, pp. 2332–2359. Available from DOI: 10.1086/500975.

23. SMEE, S. A.; GUNN, J. E.; UOMOTO, A.; *ET AL.* The Multi-Object, Fiber-Fed Spectrographs for the Sloan Digital Sky Survey and the Baryon Oscillation Spectroscopic Survey. *Astron. J.* 2013, vol. 146, no. 2, p. 32. Available from DOI: 10.1088/0004-6256/146/2/32.

24. VAN DER MAATEN, Laurens; HINTON, Geoffrey. Visualizing data using t-SNE. *Journal of machine learning research*. 2008, vol. 9, no. 11.

25. MCINNES, Leland; HEALY, John; MELVILLE, James. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*. 2018.

26. JOLLIFFE, Ian T. *Principal component analysis for special types of data*. Springer, 2002.

27. RUSTAMOV, Farukh. *SDSS/Dimensionality reduction/combined.ipynb · main · Farukh Rustamov / Astronomical_Data_ML · GitLab* [online]. 2025. Available also from: https://gitlab.fit.cvut.cz/rustafar/astronomical_data_ml/-/blob/main/SDSS/Dimensionality%20reduction/combined.ipynb. [accessed 2025-04-23].

28. BIRD, Jordan J. *Scene Classification: Images and Audio* [online]. 2020. Available also from: https://www.kaggle.com/datasets/birdy654/scene-classification-images-and-audio/. [accessed 2025-04-21].

29. RUSTAMOV, Farukh. *Jupyter Notebook: ≪scene≫* [online]. 2025. [accessed 2025-04-22].

30. HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome H; FRIEDMAN, Jerome H. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

31. SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 2014.

32. KE, Guolin; MENG, Qi; FINLEY, Thomas; WANG, Taifeng; CHEN, Wei; MA, Weidong; YE, Qiwei; LIU, Tie-Yan. Lightgbm: A highly efficient gradient boosting decision tree. In: 2017, vol. 30.

33. KOHAVI, Ron et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Ijcai*. Montreal, Canada, 1995, vol. 14, pp. 1137–1145. No. 2.

34. PEDREGOSA, Fabian; VAROQUAUX, Gaël; GRAMFORT, Alexandre; MICHEL, Vincent; THIRION, Bertrand; GRISEL, Olivier; BLONDEL, Mathieu; PRETTENHOFER, Peter; WEISS, Ron; DUBOURG, Vincent, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*. 2011, vol. 12, pp. 2825–2830.

35. KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*. 2012, vol. 25.

36. PEDREGOSA, Fabian; VAROQUAUX, Gaël; GRAMFORT, Alexandre; MICHEL, Vincent; THIRION, Bertrand; GRISEL, Olivier; BLONDEL, Mathieu; PRETTENHOFER, Peter; WEISS, Ron; DUBOURG, Vincent, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*. 2011, vol. 12, pp. 2825–2830.

37. IVEZIĆ, Željko; CONNOLLY, Andrew J; VANDERPLAS, Jacob T; GRAY, Alexander. *Statistics, data mining, and machine learning in astronomy: a practical Python guide for the analysis of survey data*. Vol. 8. Princeton University Press, 2020.

38. DIETTERICH, Thomas G. Ensemble methods in machine learning. In: *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

39. SMITH, Leslie N. Cyclical learning rates for training neural networks. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2017, pp. 464–472.

40. PRECHELT, Lutz. Early stopping—but when? In: *Neural Networks: Tricks of the Trade*. Springer, 1998, pp. 55–69.

41. FRIEDMAN, Jerome. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*. 2001, vol. 29, no. 5, pp. 1189–1232.

42. ROUSSEEUW, Peter J; CROUX, Christophe. Alternatives to the median absolute deviation. *Journal of the American Statistical association*. 1993, vol. 88, no. 424, pp. 1273–1283.

43. BALTRUSAITIS, Tadas; AHUJA, Chaitanya; MORENCY, Louis-Philippe. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2018, vol. 41, no. 2, pp. 423–443.

44. GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron; BENGIO, Yoshua. Deep learning. 2016, vol. 1, no. 2.

45. SRIVASTAVA, Nitish; HINTON, Geoffrey; KRIZHEVSKY, Alex; SUTSKEVER, Ilya; SALAKHUTDINOV, Ruslan. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*. 2014, vol. 15, no. 1, pp. 1929–1958.

46. PRECHELT, Lutz. Early stopping-but when? In: *Neural Networks: Tricks of the trade*. Springer, 2002, pp. 55–69.

47. *SDSS/ML qualities/DT_qualities.ipynb · main · Farukh Rustamov / Astronomical_Data_ML · GitLab* [online]. 2025. Available also from: https://gitlab.fit.cvut.cz/rustafar/astronomical_data_ml/-/blob/main/SDSS/ML%20qualities/DT_qualities.ipynb?ref_type=heads. [accessed 2025-04-23].

48. RUSTAMOV, Farukh. *SDSS/ML qualities/VGGNet12_qualities.ipynb · main · Farukh Rustamov / Astronomical_Data_ML · GitLab* [online]. 2025. Available also from: https://gitlab.fit.cvut.cz/rustafar/astronomical_data_ml/-/blob/main/SDSS/ML%20qualities/VGGNet12_qualities.ipynb?ref_type=heads. [accessed 2025-04-23].

49. *SDSS/ML qualities/LGBM_qualities.ipynb · main · Farukh Rustamov / Astronomical_Data_ML · GitLab* [online]. 2025. Available also from: https://gitlab.fit.cvut.cz/rustafar/astronomical_data_ml/-/blob/main/SDSS/ML%20qualities/LGBM_qualities.ipynb?ref_type=heads. [accessed 2025-04-23].

50. SAVITZKY, Abraham; GOLAY, Marcel JE. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*. 1964, vol. 36, no. 8, pp. 1627–1639.

51. ZHAO, Fei; ZHANG, Chengcui; GENG, Baocheng. Deep multimodal data fusion. *ACM computing surveys*. 2024, vol. 56, no. 9, pp. 1–36.

52. DIELEMAN, Sander; WILLETT, Kyle W; DAMBRE, Joni. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly notices of the royal astronomical society*. 2015, vol. 450, no. 2, pp. 1441–1459.