EXPERIMENT WITH MACHINE LEARNING ON HIERARCHICAL MULTI-MODAL ASTRONOMICAL DATA

Farukh Rustamov

Bachelor's thesis
Faculty of Information Technology
Czech Technical University in Prague
Department of Applied Mathematics
Study program: Applied Mathematics
Specialisation: Machine Learning and Astrophysics
Supervisor: RNDr. Petr Škoda, CSc.
May 16, 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.

Citation of this thesis: Rustamov Farukh. *Experiment with Machine Learning on Hierarchical Multi-Modal Astronomical Data*. Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2025.

Acknowledgments

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.

I further acknowledge the publicly released photometric and spectroscopic data from SDSS, without which the analysis presented here would not have been possible. Funding for the Sloan Digital Sky Survey has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org.

Finally, I thank the HiSS-Cube pipeline and its creator, Ing. Jiří Nádvorník, Ph.D., 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.

I declare that I have used AI tools during the preparation and writing of my thesis. I have verified the generated content. I confirm that I am aware that I am fully responsible for the content of the thesis.

In Prague on May 16, 2025

Abstract

This thesis presents a comprehensive study of the prediction of star formation rates (SFR) in galaxies using multimodal data acquired from the Sloan Digital Sky Survey (SDSS, DR7). We first filter the initial SFR catalogue to produce a high-quality, cleaned subset of 11,179 galaxies that have valid one-dimensional spectra and five-band photometry. This subset is then run through the HiSS-Cube pipeline to generate various image resolutions ranging from 64×64 to 8×8 pixels and various samplings of spectral data ranging from 4620 to 577 bins while retaining corresponding measurement uncertainties.

Three groups of regression models were considered in this thesis: Decision Tree (DT), the VGGNet12 convolutional neural network, and LightGBM gradient boosting, for three different modalities: photometry-only, spectroscopy-only, and multimodal fusion, including both early and late fusion. The hyperparameter search is conducted via grid search with five-fold cross-validation, and model performance assessment is done via a range of metrics, such as R^2, MAE, RMSE a NMAD.

The top performance is obtained by the early-fusion LightGBM model with $R^2 = 0.308$, MAE=0.19, RMSE=0.32, illustrating the power of tree-based learners for fusing visual and spectral features. VGGNet12 trained on photometric images alone also performs well ($R^2 = 0.262$), illustrating the power of deep CNNs for morphological feature learning. Interestingly, the lowest spectral resolution generalizes better because of implicit noise smoothing.

Our results validate that multimodal machine learning can record complementary astrophysical features for SFR estimation accurately. Methodological groundwork laid here allows for the exploration of more sophisticated fusion methods and utilization of other data modalities incorporating the information about the circumgalactic environment or galaxy kinematics in the future.

Keywords machine learning, SDSS, star formation rate, spectroscopy, photometry, multimodal fusion, HiSS-Cube

Abstrakt

Tato bakalářská práce představuje komplexní studii predikce rychlosti tvorby hvězd (SFR) v galaxiích pomocí multimodálních data získaných z přehlídky oblohy Sloan Digital Sky Survey (SDSS, DR7). Nejprve filtrujeme počáteční katalog SFR, abychom vytvořili vysoce kvalitní, vyčištěnou podmnožinu 11,179 galaxií, které mají platná jednorozměrná spektra a fotometrii (images) v pěti pásmech. Tato podmnožina je následně zpracována pomocí pipeline HiSS-Cube, která generuje různé obrazové rozlišení od 64×64 do 8×8 pixelů a různá vzorkování spektrálních dat od 4620 do 577 binů při zachování odpovídajících měřicích nejistot.

V této práci jsou použity tři skupiny regresních modelů: rozhodovací strom, konvoluční neuronovou síť VGGNet12 a gradientní boosting LightGBM, pro tři různé modality: pouze fotometrie, pouze spektroskopie a multimodální fúze, včetně časné a pozdní fúze. Hledání hyperparametrů je prováděno pomocí metody grid search s pětinásobnou křížovou validací a hodnocení výkonu modelu je provedeno pomocí různých metrik, jako jsou R^2, MAE, RMSE a NMAD.

Nejlepšího výkonu dosahuje model LightGBM s časnou fúzí s hodnotami $R^2 = 0{,}308$, MAE=0,19, RMSE=0,32, což ilustruje sílu stromových algoritmů při slučování vizuálních a spektrálních znaků. VGGNet12 trénovaná pouze na fotometrických obrazech také dosahuje dobrých výsledků ($R^2 = 0{,}262$), což ukazuje sílu hlubokých CNN pro učení morfologických rysů. Zajímavé je, že nižší spektrální rozlišení generalizuje lépe díky implicitnímu vyhlazování šumu.

Naše výsledky potvrzují, že multimodální strojové učení dokáže přesně zachytit komplementární astrofyzikální příznaky pro odhad SFR. Metodologický základ položený v této práci umožňuje v budoucnu zkoumat sofistikovanější metody fúze a využívat další datové modality zahrnující informaci o okolním prostředí či kinematice dané galaxie.

Klíčová slova strojové učení, SDSS, rychlost formování hvězd, spektroskopie, fotometrie, multimodální fúze, HiSS-Cube.

Contents

List of Figures

List of Tables

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
LLM	Large Language Model
LightGBM	Light Gradient Boosting Machine
HDF5	Hierarchical Data Format version 5
RCI	Research Computing Infrastructure
PCA	Principal Component Analysis
t-SNE	t-Distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
VO	Virtual Observatory
HiSS-Cube	Hierarchical Semi-Sparse Cube
API	Application Programming Interface
FITS	Flexible Image Transport System
HEALPix	Hierarchical Equal Area isoLatitude Pixelation
GPU	Graphics Processing Unit
SLURM	Simple Linux Utility for Resource Management
CPU	Central Processing Unit
FWHM	Full Width at Half Maximum
AGN	Active Galactic Nucleus
ZTF	Zwicky Transient Facility
LSST	Legacy Survey of Space and Time
APOGEE	Apache Point Observatory Galactic Evolution Experiment
LAMOST	Large Sky Area Multi-Object Fiber Spectroscopic Telescope
AGB	Asymptotic Giant Branch
ISM	Interstellar Medium
CCD	Charge-Coupled Device
GOSS	Gradient-based One-Side Sampling
EFB	Exclusive Feature Bundling

Chapter 1

Introduction

1.1 General Description and Relevance of the Study

Multimodal machine learning has witnessed remarkable advancements in recent years.

The field of study has applications extending from independent driving and medical diagnostics into the analysis of astronomical data.

The combination of various data forms—such as images, text, audio, and structured signals—enables models to learn more subtle representations and make more accurate forecasts in complex settings.

In astrophysics, large-area surveys like the Sloan Digital Sky Survey (SDSS) [1] offer both photometric and spectroscopic information for millions of astrophysical sources.

These synergistic modalities offer unique insights: images capture structural and morphological characteristics, whereas spectra hold detailed physical and chemical properties.

This thesis focuses on the applications of multimodal machine learning techniques to predict the star formation rate (SFR) [2] in galaxies using data from SDSS.

The need is motivated by the requirement to process gigantic astronomical datasets and construct models that take advantage of the strengths of both image-based and spectroscopic input.

1.2 HiSS-Cube Software Infrastructure

A wide variety of approaches exist for visualizing and analyzing large astronomical data cubes, but most either rely on static FITS files or lose the native measurement uncertainties when building coarser resolutions. To address these limitations, Ing. Jiří Nádvorník, Ph.D. developed the *Hierarchical Semi-Sparse Cube (HiSS-Cube)* framework [3] based on HDF5, which offers:

- **Multi-domain fusion:** Supports imaging, spectral, environmental and time-series data in a single hierarchical cube.
- **Preserved uncertainties:** Constructs lower-resolution representations without discarding per-pixel or per-bin error estimates.
- **Scalability:** Leverages hierarchical indexing (HEALPix) and semi-sparse storage to enable rapid spatial queries over billions of measurements.
- **Machine-learning ready:** Exports arbitrary resolution cutouts to contiguous NumPy arrays, avoiding repeated I/O or reprocessing when exploring different model input sizes.
- **Virtual Observatory compatibility:** Exports to VOTable/FITS for use in standard VO tools.
- **Performance gains:** Benchmarks on SDSS Stripe 82 show HiSS-Cube queries are orders of magnitude faster than raw FITS exports for both interactive visualization and large-scale ML pipelines.

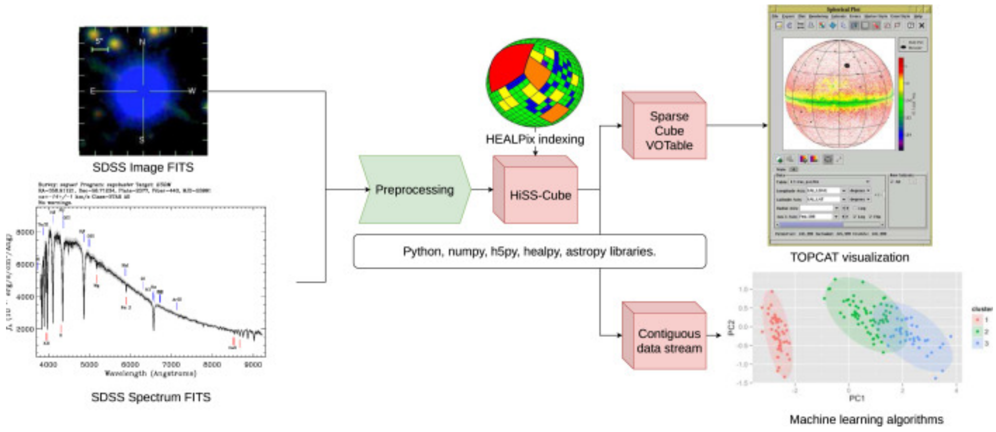


Figure 1.1 HiSS-Cube data-flow pipeline: from SDSS raw FITS files to multi-layered, semi-sparse HDF5 cubes for visualization and machine learning. Image from [3].

The core idea is to precompute a hierarchy of semi-sparse, multi-resolution cubes that retain scientific uncertainties at every scale. This allows, for example, an ML workflow to first coarse-scan a large region, then seamlessly drill down to higher resolutions without re-ingesting or re-calibrating the data.

Data Lineage and Target Sample. Our analysis is based on the MPA-JHU value-added catalog [4], in which star formation rates (SFRs) were computed for galaxies from the Sloan Digital Sky Survey (SDSS) using the method described by Brinchmann et al. (2004). This catalog provides a table containing galaxy identifiers and their corresponding computed SFR values.

Using the HiSS-Cube [3] infrastructure deployed on the Karolina supercomputing cluster in Ostrava, previous researchers executed complex queries on the complete original SDSS dataset (approximately 50 TB of images and spectra) to extract images and spectra corresponding to the galaxies listed in this catalog. These data were then preprocessed by generating multiple resolution versions of the images and spectra and calculating their associated measurement uncertainties. This process resulted in the HiSS-Cube dataset of approximately 4 TB, stored in a hierarchical semi-sparse cube format (HDF5). Subsequently, this prepared dataset was transferred from Ostrava to the RCI computational cluster at the Faculty of Electrical Engineering, Czech Technical University in Prague [5].

In this work, we performed all subsequent analyses—including data filtering, multimodal fusion modeling, and regression model benchmarking—using the RCI cluster.

Multimodal application. One concrete use case demonstrated here is the end-to-end SFR regression pipeline: from a single API call we retrieve both multi-band image filters and one-dimensional spectral vectors for each galaxy, fully preserving uncertainties and spatial indexing. This seamless integration underlies the early- and late-fusion experiments detailed in Chapters 3 and 4, and illustrates how HiSS-Cube can accelerate the development and deployment of advanced multimodal machine-learning workflows in astronomy.

Software availability. The complete HiSS-Cube codebase—including source, documentation, and issue tracker—is openly maintained on GitHub [6].

1.3 Computational Environment: RCI Cluster

All large-scale data processing and model training were performed on the RCI [5] (Research Computing Infrastructure) cluster at CTU-FEL, since the multi-terabyte HiSS-Cube datasets and deep learning workloads are not feasible on a local workstation. Our jobs were submitted via SLURM to the `gpufast` partition with the following resource request:

- 1 GPU, 8 CPU cores, and 128 GB RAM
- a hard wall-time limit of 4 hours on the GPU node
- access to the shared parallel filesystem at `/mnt/data`

Environment management was handled via Miniconda, using the `myenv` environment for all Python dependencies.

For interactive work, we launched Jupyter Notebook on the compute node and tunneled it to a local machine.

All experiments—including data ingest, preprocessing, hyperparameter sweeps, and fusion model training—ran under this 4 h GPU limit, enabling rapid iteration on large-scale astronomical datasets that would be impractical on a desktop machine.

1.4 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. [1], and DR7 represents the completion of the Legacy Survey, covering over 8,000 deg^2 with more than 1.6 million galaxy spectra [7]. 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 [8].

In this thesis we primarily use data from SDSS Data Release 7 (DR7) [9], because the star formation rates we employ were computed for that release. 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.

1.4.1 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 filters.
- **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 Spectroscopy vs. Photometry Role

Spectroscopic data offer direct physical diagnostics—emission line luminosities (e.g., Hα), which are tied to the instantaneous star formation rate (SFR), and redshift estimates for distance corrections [10]. Photometric images capture morphological details, color gradients, and total broadband flux, echoing the stellar content and dust properties of the galaxy. By merging these complementary perspectives, our models are able to take advantage of both fine-grained spectral physics and large-scale structural indicators, resulting in enhanced and more precise SFR estimates.

1.5 Research Challenges

Working with the SDSS data presents several challenges:

1. **Data Filtering.** The original SDSS SFR catalog has over 4.8 million records, but only a small proportion have both good multi-band filters and valid SFR measures. Objects with missing photometry or spectroscopy, undefined SFR values (designated as NaN or the placeholder value -99), and non-galactic objects must be excluded, reducing the dataset to a few thousand galaxies amenable to regression analysis [11].
2. **Quality of Images and Spectra.** The HiSS-Cube pipeline provides four image resolutions (64×64, 32×32, 16×16, 8×8 px) and four spectral samplings (4620, 2310, 1155, 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 [3].
3. **Multiple Objects within a Single Image.** SDSS cutouts may contain overlapping stars or galaxies, resulting in blended light profiles that mislead subsequent feature extractors. To prevent each input being a collection of multiple target galaxies, we use automatic segmentation using thresholding and connected-component labeling, marking and discarding multi-object cutouts [12, 13].

1.6 Aims and Responsibilities

The overall aim of this thesis is to come up with the best practice in SFR prediction from SDSS data. To achieve this goal, the following tasks will be completed:

1. Perform a stringent examination of raw data, assess its quality, and apply filtering.

2. Develop algorithms to aim at the automatic identification and separation of objects in images.
3. Investigate the effect of variations in image and spectrum quality on prediction accuracy.
4. Contrast the performance of single modality models with multimodal approaches.
5. Perform a comparative analysis of publicly accessible Scene dataset, calibrating the results to match SDSS to guarantee our multimodal pipeline under controlled conditions.
6. Quantify the relative performance gain of multimodal fusion compared to unimodal (image-only and spectrum-only) baselines on a structurally comparable external test set in order to demonstrate added value by combining modalities.
7. Compare and analyze the training and inference times of all models and modalities on both the SDSS and the external dataset to investigate 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 [14].

1.7.1.2 The Rationale and Significance of Spectral Analysis

- **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 β lines is a common metallicity diagnostic [15]. These abundance measurements are crucial for understanding galactic chemical evolution and enrichment histories [14].

- **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 [15]. 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 [14]. Understanding these conditions informs models of star-formation efficiency and feedback processes.

All of these diagnostics are discussed in [14, p. 1–6].

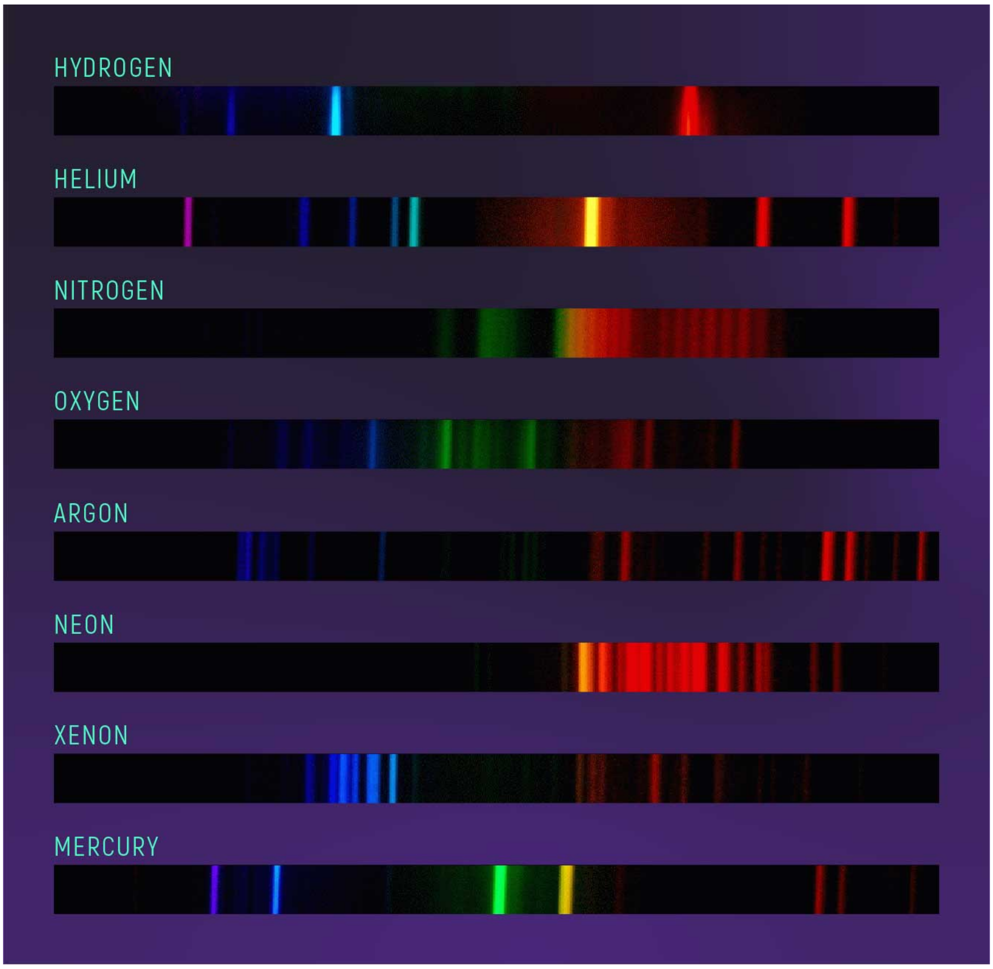


Figure 1.2 Example of atomic spectral lines for different elements.[16]

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 $z = 913\,\text{nm}$ as shown in Figure 1.3. 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 $\Delta z \approx 95\,\text{nm}$ [17].

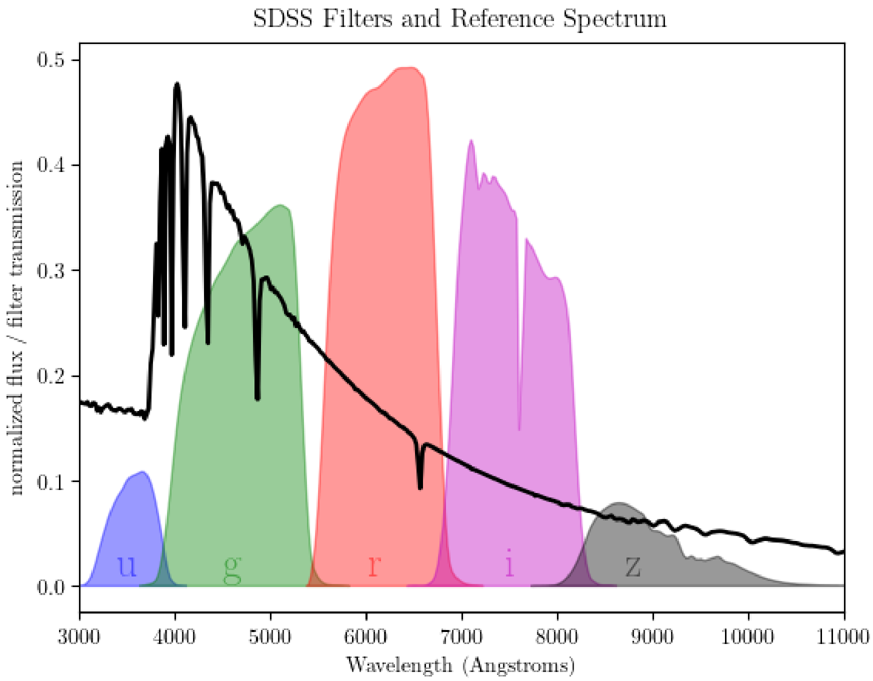


Figure 1.3 Transmission curves of the SDSS u, g, r, i, z filters.

1.7.3 Star Formation Rate (SFR)

1.7.3.1 Definition and Conceptual Scope of the Star Formation Rate (SFR)

The star formation rate (SFR) measures how quickly a galaxy turns its available gas into new stars. It is given in solar masses per year ($M_\odot\,\mathrm{yr}^{-1}$), meaning, for example, that an SFR of 1 $M_\odot\,\mathrm{yr}^{-1}$ corresponds to the formation of one Sun's worth of stars each year.

Beyond describing the galaxy's current activity, SFR also helps us trace its life story: by comparing the present SFR to the average over past epochs, we can tell if the galaxy is quietly aging, steadily forming stars, or experiencing a starburst. This comparison uses the *birthrate parameter*

$$b = \frac{\mathrm{SFR}_{\mathrm{current}}}{\langle \mathrm{SFR}_{\mathrm{past}} \rangle},$$

where $b < 1$ indicates a slowdown, $b \approx 1$ steady formation, and $b > 1$ a recent burst of star formation [18]. On cosmic scales, the average SFR density rose to a peak around redshift $z \sim 2$ (about 10 billion years ago) and has since declined by an order of magnitude [19].

1.7.3.2 Star Formation Rate (SFR) as a Fundamental Parameter of Galaxies

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 [**KennicuttEvans2012**].
- **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 [20].
- **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 [21].
- **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 [22].
- **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 [19].

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 [4]. To ensure that every galaxy has both imaging and spectroscopic data, we retain only those entries with available multi-band filters 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`≠0 [11, 23]. 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`≠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 four image quality levels and four spectral resolutions [3]. Moreover, each of these variants carries the same `AVG` SFR label, simplifying our supervised learning setup.

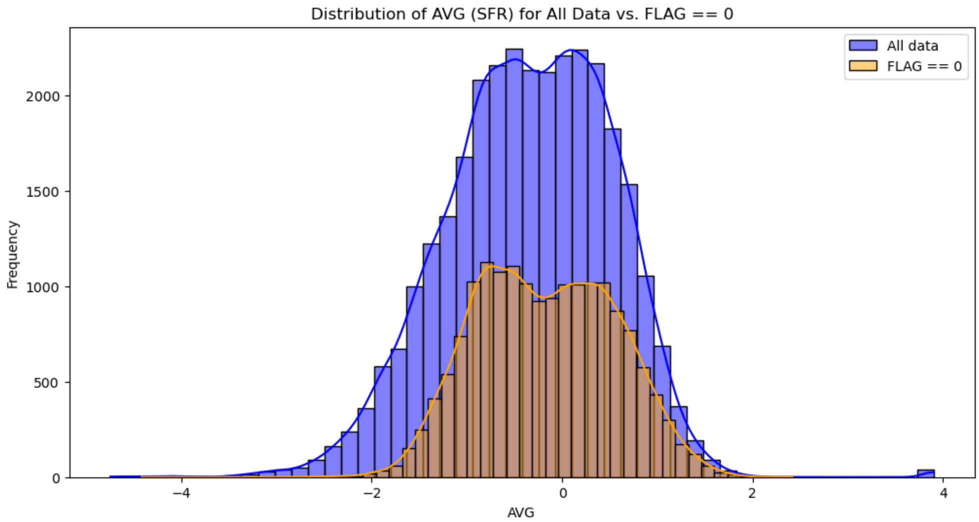


Figure 2.1 Distribution of `AVG` ($\log_{10}$ SFR) in the filtered sample.

Figure 2.1 reveals a roughly log-normal distribution of SFR values, with most galaxies clustered around $\log_{10}(\text{SFR}) \sim -1.5$ to 1.5.

2.2 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) [17]. 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 ob-

ject is provided with a spectrum that offers information on its chemical composition, temperature, and dynamics.

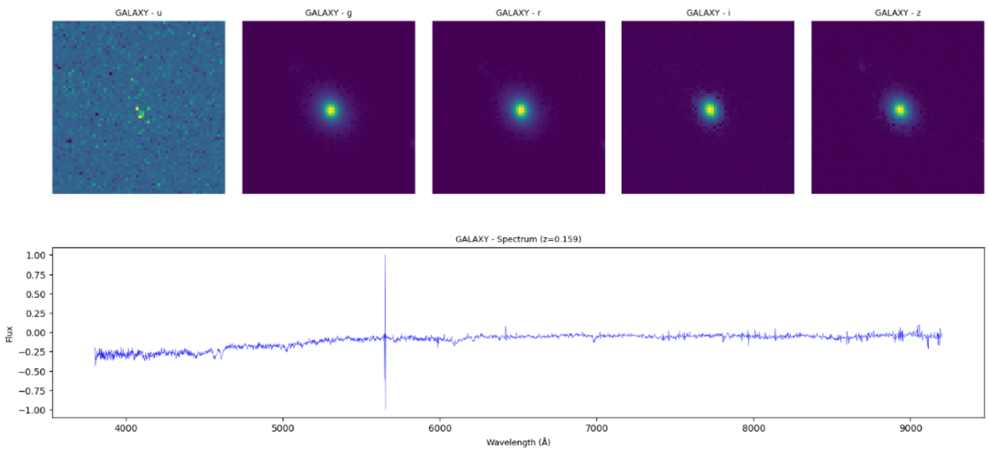


Figure 2.2 An example of the SDSS galaxy. At the top: five filter images; at the bottom: a spectrum.

2.3 Image and Spectrum Data Availability

Thanks to the HiSS-Cube pipeline [3], each high-quality galaxy (`FLAG`=0) is preprocessed into a multi-resolution "cube" that preserves uncertainties. For our regression experiments, we retrieve four image resolutions and four spectral samplings per object.

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

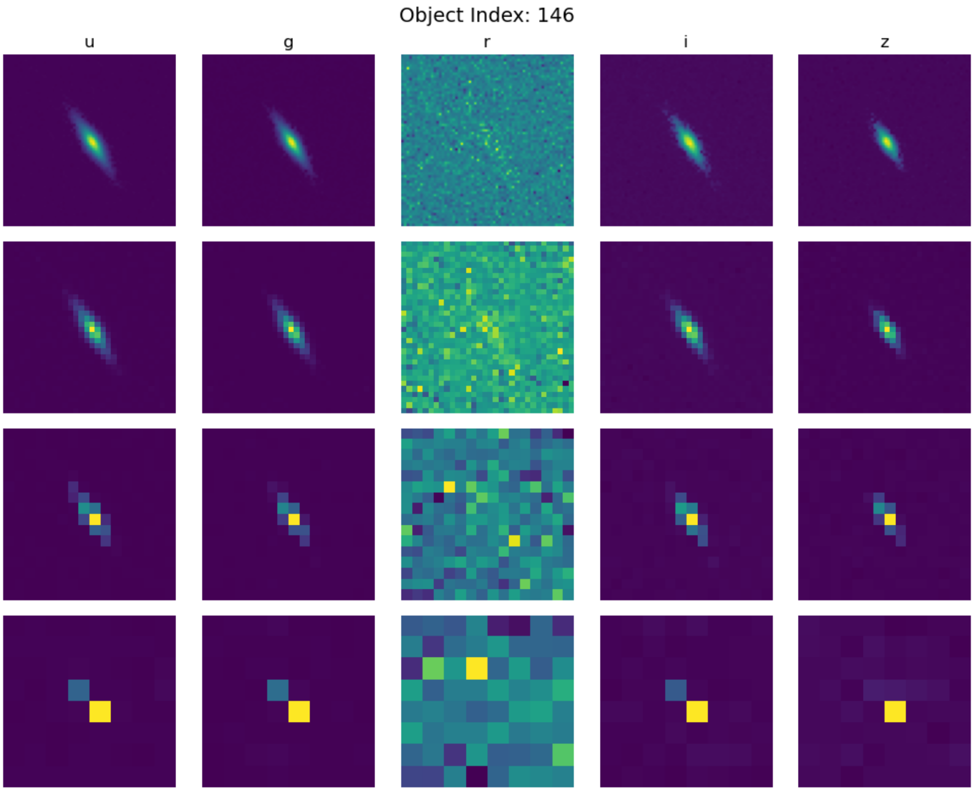


Figure 2.3 HiSS-Cube image outputs for a single galaxy at four resolution levels (64×64 to 8×8 pixels).

- **Spectral vectors.** Four one-dimensional samplings with length $L \in \{4620, 2310, 1155, 577\}$ bins, obtained by uniform downsampling of the native SDSS spectrum. Lower-resolution spectra effectively smooth high-frequency noise, serving as a built-in denoiser.

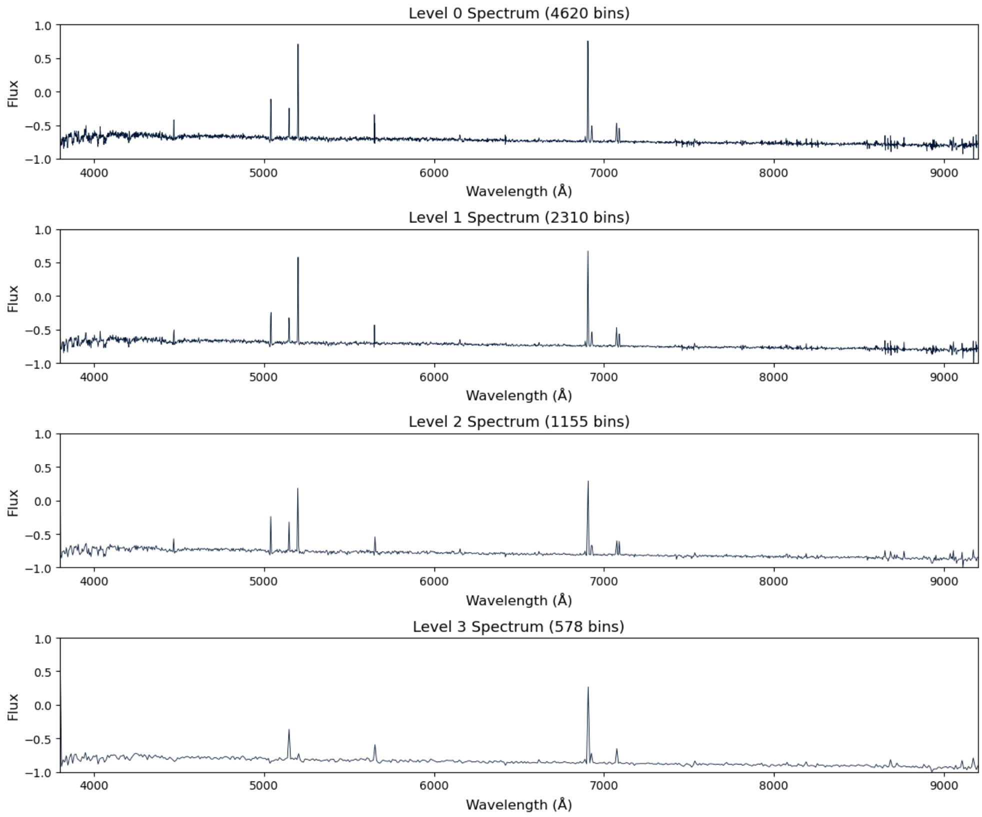


Figure 2.4 HiSS-Cube spectral outputs for the same galaxy at four sampling levels (4620 to 577 bins).

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

2.4 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.” [11]

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

2.5 Analysis of NaN Block Lengths and Positions

2.5.1 NaN Percentage by Object

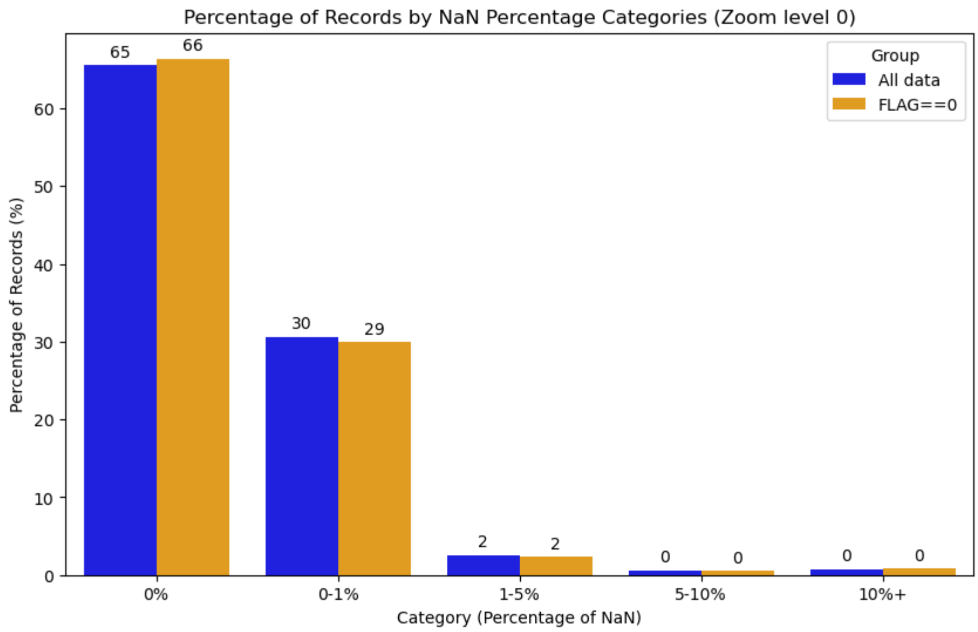


Figure 2.5 Percentage of records by NaN percentage categories at Zoom level 0, comparing all data vs. FLAG=0 subset.

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.

2.5.2 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.

2.5.3 Distribution of NaN Run Lengths

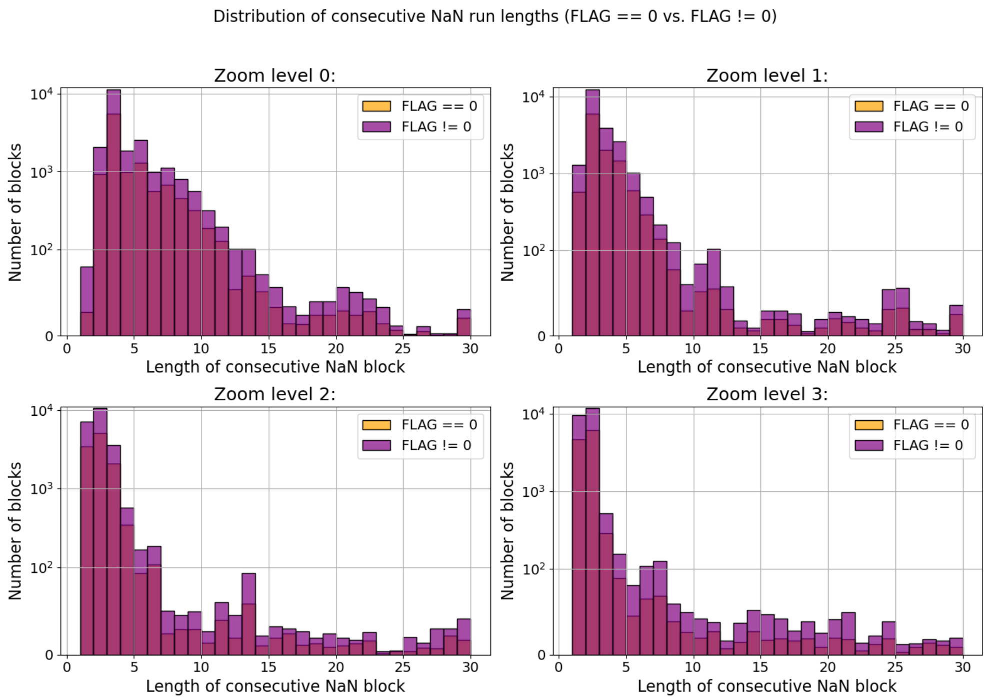


Figure 2.6 Distribution of consecutive NaN run lengths at each resolution for `FLAG`=0.

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; however, there are exceptions—large spectral gaps do occur, for example as shown in Figure 2.8(a).

2.5.4 NaN Occurrence Along Wavelength

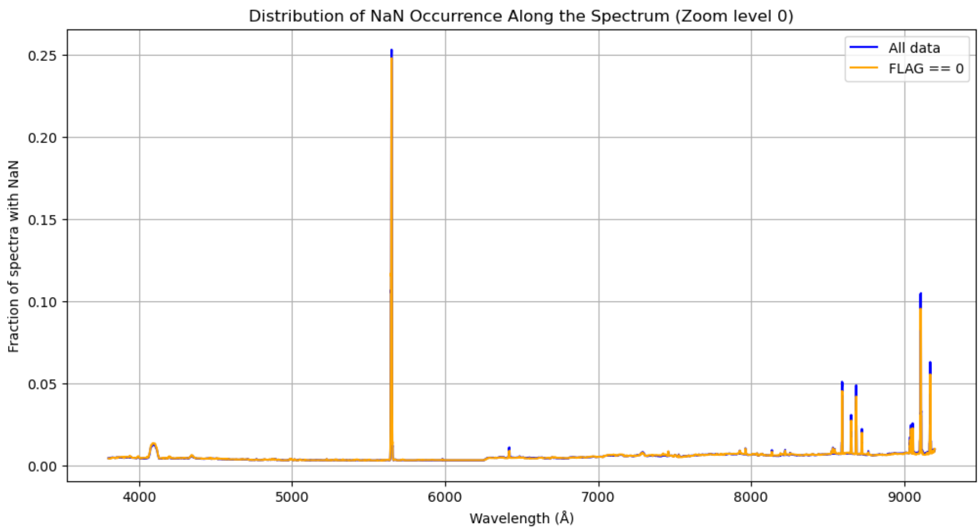


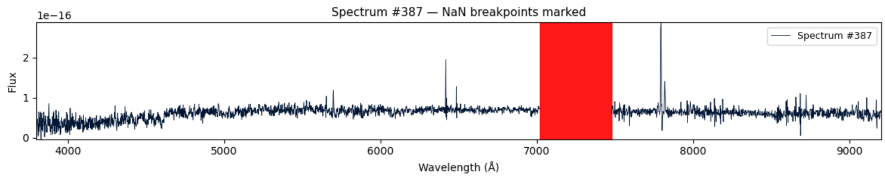
Figure 2.7 Typical wavelength regions where NaN gaps commonly occur (Zoom level 0).

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.

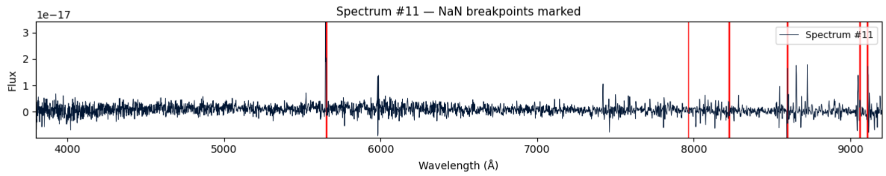
Each point along the wavelength axis represents the fraction of spectra in which that specific bin is flagged as NaN. Noticeably, there is no wavelength where 0% of spectra are missing data. It indicates that every channel is affected by occasional dropouts or quality flags. The sharp spike at ~ 5500 Å coincides with the dichroic split between the blue and red arms of the SDSS spectrograph. At this intersection, stitching mismatches and calibration uncertainties often lead to flagged pixels [24].

The elevated NaN occurrence near ~ 9000 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 [25].



(a) SDSS Spectrum #387 showing a continuous NaN region (shaded red) from approximately 7000 Å to 7500 Å.



(b) SDSS Spectrum #11 with multiple discrete NaN gaps, marked by red vertical lines at about 5800 Å, 8000 Å, 8950 Å, and 9150 Å.

Figure 2.8 Examples of SDSS spectra containing missing (NaN) segments. Red highlights indicate the affected wavelength regions.

2.6 Detection and Removal of Multi-Object Cutouts

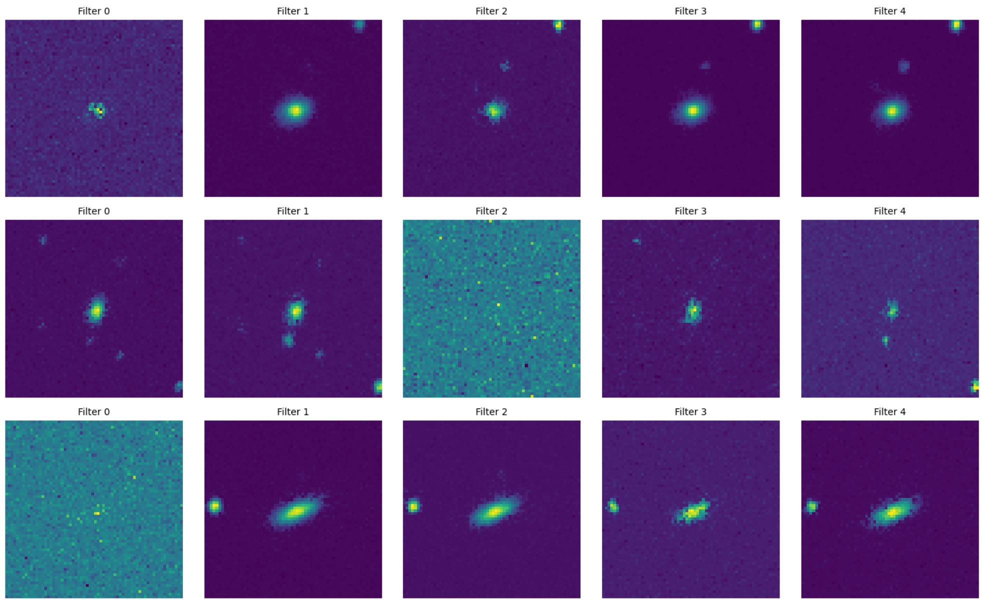


Figure 2.9 Example of a cutout containing multiple detected sources, excluded from the final sample [23] .

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 [12]. 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 [23] and closely follows the methodology of Sezgin and Sankur's survey on thresholding techniques [12] as well as the standard workflow described in Gonzalez and Woods's digital image processing text [13].

2.7 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

2.7.1 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** [26] — a nonlinear technique that preserves local structure by minimizing the Kullback–Leibler divergence between probability distributions of pointwise neighborhoods in high- and low-dimensional spaces. t-SNE first converts pairwise similarities in the high-dimensional space into joint probabilities using a Gaussian kernel, then defines analogous joint probabilities in the low-dimensional map via a Student's t-distribution to alleviate the "crowding problem." By iteratively minimizing the KL divergence through gradient descent, t-SNE excels at revealing fine-grained cluster structure and manifold substructure. Its main advantages are strong separation of local clusters and intuitive visual grouping, though it can be computationally intensive and sensitive to hyperparameters such as perplexity.
- **UMAP** [27] — a topological manifold learning algorithm that constructs a fuzzy simplicial complex in high dimensions and optimizes its low–dimensional embedding to preserve both local and global data structure. UMAP models the data manifold by estimating a weighted graph of nearest neighbors, then applies stochastic gradient descent to minimize the cross-entropy between the high-dimensional and low-dimensional fuzzy complexes. This yields embeddings that faithfully maintain global neighbor relations while still clustering similar samples tightly. Compared to t-SNE, UMAP is

typically faster on large datasets, offers greater control via explicit nearest-neighbor and minimum-distance parameters, and often produces more meaningful global layouts.

- **PCA** [28] — a linear method that identifies orthogonal directions (principal components) of maximum variance in the data and projects the data onto the leading components for dimensionality reduction. PCA computes the eigenvalues and eigenvectors of the empirical covariance matrix, ordering components by explained variance. This yields a deterministic, interpretable embedding in which each axis corresponds to a linear combination of original features. Its advantages include simplicity, scalability to very high dimensions, and the ability to capture the largest sources of variance; however, PCA cannot capture nonlinear relationships and may mix multiple underlying factors in each principal component.

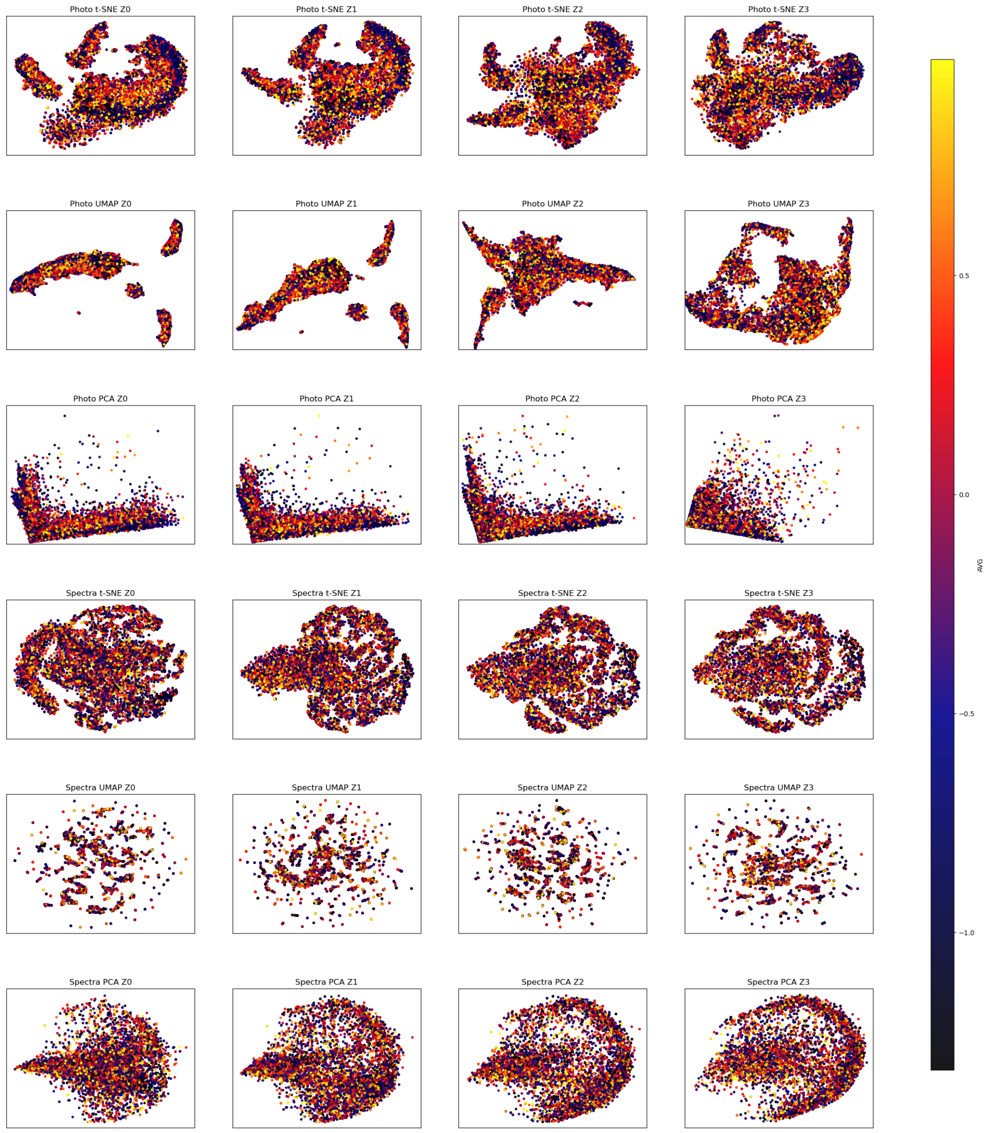


Figure 2.10 Embeddings of image and spectral data at four zoom levels (Z0–Z3) using t-SNE, UMAP, and PCA, colored by `AVG` [29].

Here, ρ_x and ρ_y are the Pearson correlation coefficients between the first (x) and second (y) embedding dimensions and the target variable `AVG` ($\log_{10}$ SFR). We computed these correlations over the full sample to assess how linearly each low-dimensional axis relates to the true SFR values [30].

Note: In what follows, Z denotes the zoom (image/spectrum) quality level:

Z0 corresponds to the highest resolution (most pixels), and Z3 to the lowest (fewest pixels).

Table 2.3 Pearson correlation coefficients (ρ_x, ρ_y) between the two embedding axes and `AVG` for each method and modality, across four zoom levels (Z0–Z3).

Method / Modality	ρ_x	ρ_y	Method / Modality	ρ_x	ρ_y
t-SNE Image Z0	-0.03	-0.04	t-SNE Spectra Z0	-0.10	+0.06
t-SNE Image Z1	-0.03	-0.06	t-SNE Spectra Z1	-0.15	+0.02
t-SNE Image Z2	0.00	-0.09	t-SNE Spectra Z2	-0.16	+0.00
t-SNE Image Z3	-0.06	-0.03	t-SNE Spectra Z3	-0.15	-0.02
UMAP Image Z0	+0.03	+0.00	UMAP Spectra Z0	+0.02	-0.02
UMAP Image Z1	+0.03	+0.03	UMAP Spectra Z1	-0.03	+0.00
UMAP Image Z2	-0.05	-0.01	UMAP Spectra Z2	-0.02	-0.01
UMAP Image Z3	+0.08	-0.02	UMAP Spectra Z3	-0.02	+0.02
PCA Image Z0	-0.03	+0.01	PCA Spectra Z0	-0.11	+0.05
PCA Image Z1	-0.05	-0.01	PCA Spectra Z1	-0.14	+0.03
PCA Image Z2	-0.07	-0.04	PCA Spectra Z2	-0.14	+0.01
PCA Image Z3	-0.07	+0.03	PCA Spectra Z3	-0.14	-0.01

The embedding analysis reveals:

- **t-SNE** and **UMAP** uncover local, nonlinear structure but show weak linear correlation with `AVG`, indicating complex manifold relationships [26, 27].
- **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 [28].

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

Multimodal Machine Learning

3.1 Introduction to Multimodal Machine Learning

Multimodal machine learning combines different types of data—like images, text, audio, and structured signals—into a single framework. It then analyzes these varied inputs together to understand how they relate and complement each other. By doing this, the resulting models often perform better than those built on just one data type [31].

In recent years, the commercial success of large language models (LLMs) has demonstrated the power of combining multiple modalities: modern systems fuse text, vision, and speech inputs to drive applications in customer service, content creation, and scientific research. For example, vision-language models enable image editing via natural-language prompts, while speech-enabled assistants interpret spoken commands in context. These successes underscore the growing importance of multimodal approaches across industries and research domains.

In this thesis, we apply multimodal learning to the astrophysical problem of predicting galaxy star formation rates (SFRs) from Sloan Digital Sky Survey (SDSS) data. The SFR regression task naturally lends itself to multimodal modeling because photometric images encode morphological structure and color information, while spectroscopic measurements trace detailed physical diagnostics such as emission-line luminosities.

3.2 Identifying Interesting Science Cases

While our primary focus is on predicting galaxy star-formation rates (SFRs), the multimodal framework developed here readily extends to a variety of other compelling astrophysical problems. Below we highlight five key science cases where fusing imaging and spectroscopic data offers significant advantages over single-modality approaches:

1. **Galaxy Morphological Classification and Evolution.** Citizen-science projects such as Galaxy Zoo have demonstrated the power of visual morphology for tracing galaxy formation pathways [32]. By combining high-resolution photometry with spectral line diagnostics (e.g. Hα/Hβ ratios), one can refine morphological classes (spiral, elliptical, irregular) and link them quantitatively to stellar population ages and dust content [33]. Multimodal models can thus map the 'Hubble sequence' onto physical parameters, uncovering subtler evolutionary trends than either modality alone can reveal.

2. **Rare Object and Anomaly Detection.** Identifying quasars, strong gravitational lenses, or low-metallicity dwarfs requires scanning through millions of sources with imbalanced class frequencies. Imaging alone often struggles with line-of-sight blends, while spectroscopy alone misses morphological context. Multimodal classification has been shown to boost purity and completeness in quasar selection [34] and to discover new strong-lens candidates by correlating arc-like features with emission-line redshift discrepancies [35, 36]. Similarly, anomaly-detection pipelines trained on both modalities can flag novel astrophysical events for follow-up [37].

3. **Transient and Variable Source Characterization.** Time-domain surveys (e.g. ZTF, LSST) deliver light curves that capture the photometric variability of supernovae, tidal disruption events, and active galactic nuclei (AGN). When spectral snapshots are also available, fusing temporal, photometric, and spectroscopic features enables more accurate classification of transients [37, 38]. For example, embedding a supernova's spectral line velocities alongside rise-time photometry has improved subtype separation (Ia vs. IIn) and can accelerate spectroscopic follow-up decisions.

4. **Stellar Parameter Inference and Peculiar Star Identification.** Large-scale stellar surveys (e.g. APOGEE, LAMOST) provide both multi-band imaging and high-resolution spectra. Jointly modeling a star's color–-magnitude position with its detailed absorption-line profile allows for more precise determinations of effective temperature, surface gravity, and metallicity [39, 40]. Moreover, multimodal outlier detection has uncovered rare stellar populations—such as carbon stars and peculiar white dwarfs—by high-

lighting discrepancies between photometric and spectroscopic parameter estimates.

5. **Environmental Effects on Galaxy Properties.** The interplay between a galaxy's local density (cluster vs. void) and its internal processes drives quenching and morphological transformation. By fusing imaging (tracing morphology and tidal features), spectroscopy (tracing emission-line strengths and kinematics), and environmental metrics (e.g. 5th-nearest-neighbor density), one can disentangle competing mechanisms such as ram-pressure stripping vs. galaxy harassment [41, 42]. Multimodal regression models can quantify how environment modulates SFR beyond global scaling relations, providing a path to unravel the drivers of cosmic star-formation history.

In each of these cases, the complementary strengths of photometry (morphology, spatial context) and spectroscopy (physical diagnostics, redshift precision) combine to yield richer, more robust scientific inferences than single-modality analyses. Of course, this adaptation also requires assembling a dedicated team with the appropriate expertise and carefully considering the statistical distribution of the data for the problem at hand. HiSS-Cube pipeline and our multimodal architectures can be readily adapted to these problems by swapping the SFR label for the relevant target (e.g. morphology class, stellar parameters, transient type) and retraining under the same fusion paradigms.

Another popular science case where multimodal machine learning can help is the classification of stars, galaxies, and quasars. 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. Here, the dataset refers to the version we preprocessed and explored in Chapter 2 (Data Exploration).

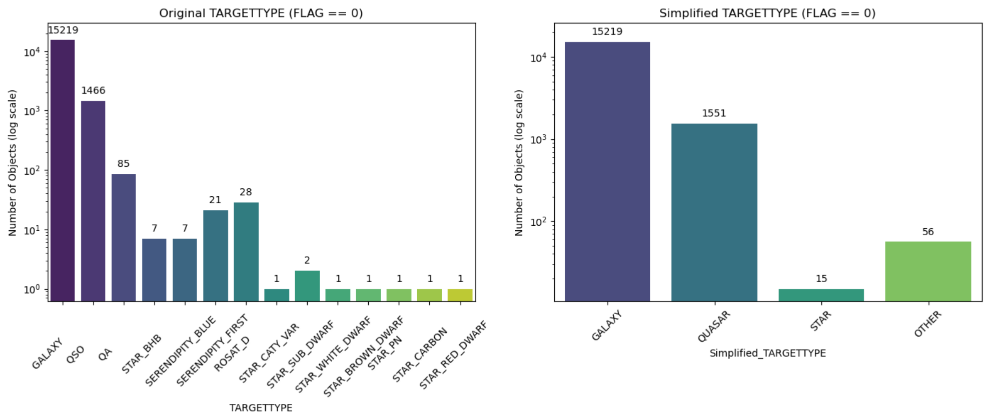


Figure 3.1 Class distribution for star–galaxy–quasar labels: galaxies outnumber quasars by a factor of 10, and stars comprise fewer than 30 objects [23].

3.3 Fusion Strategies in Multimodal Learning

A key design choice in multimodal systems is how and when to combine information from different modalities. Two canonical approaches are:

Early Fusion Feature-level fusion where modality-specific features are extracted independently and then concatenated (or otherwise merged) into a joint embedding, which is passed to a single model for prediction. Early fusion enables cross-modal feature interactions from the very beginning of the learning process. [43]

Late Fusion Decision-level fusion where each modality is processed by its own model, producing independent predictions, which are then combined (e.g., averaged or weighted) to yield the final output. Late fusion simplifies model training by decoupling modality-specific learners and often improves robustness by enforcing model diversity. [43]

3.4 Scene Dataset Example

To illustrate the general benefits of multimodal learning, we conducted preliminary experiments on the publicly available Scene dataset [44], which provides two modalities for environmental scene classification:

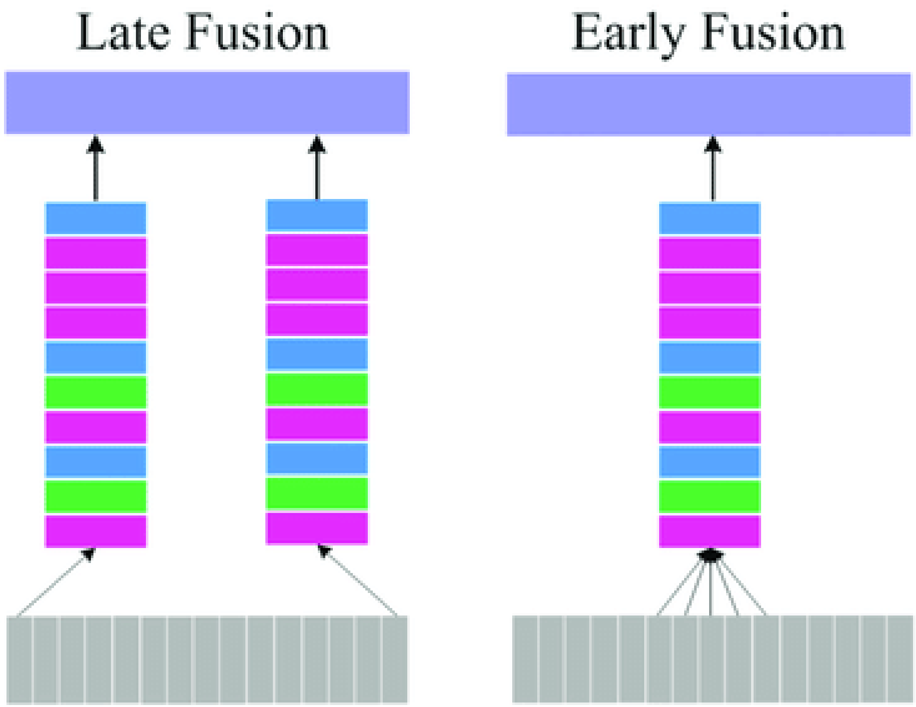


Figure 3.2 Illustration of Late and Early Fusion strategies in multimodal learning [43].

- **Images:** Still frames depicting eight scene types (e.g., beach, classroom, forest).
- **Audio Features:** Mel-Frequency Cepstral Coefficients (MFCCs) extracted from audio recordings synchronized with each image.

The dataset supports two hierarchical classification tasks:

- **CLASS1:** Binary classification of scenes as indoors vs. outdoors.
- **CLASS2:** Fine-grained classification into eight specific scene categories.

3.4.1 MFCC Features.

Mel-Frequency Cepstral Coefficients (MFCCs) are compact representations of the short-time power spectrum of an audio signal, mapped onto the Mel scale

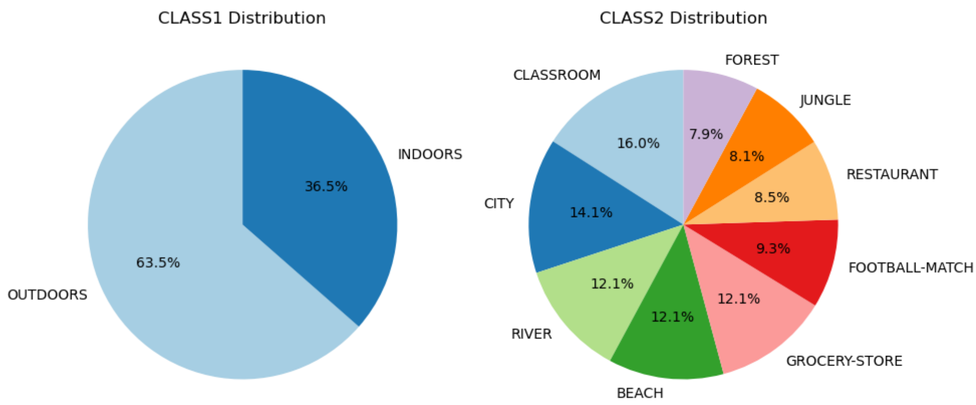


Figure 3.3 CLASS1 (left) and CLASS2 (right) label distributions for the Scene dataset.

to mimic human auditory perception. By extracting the first 13–40 MFCCs (plus their first and second derivatives) from each synchronized audio clip, we obtain low-dimensional feature vectors that capture timbral and spectral patterns—rather than using raw waveform samples. All of our audio-based models are therefore trained on these MFCC representations, which improves robustness to noise and drastically reduces input dimensionality. [45]

3.4.2 Qualitative Example of Image–Audio Pair

To further illustrate the complementarity of modalities, Figure 3.4 shows a representative scene image alongside its Mel-Frequency Cepstral Coefficients spectrogram extracted from the synchronized audio clip.

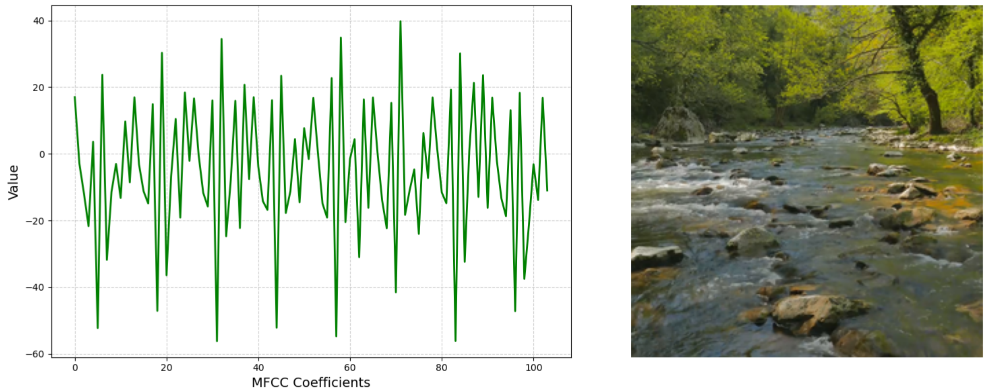


Figure 3.4 Qualitative example of the multimodal input: the visual scene and its audio features.

Quantitative Results

Table 3.1 summarizes the CLASS1 and CLASS2 test accuracies for a Decision Tree classifier (max_depth=12, class_weight=balanced) trained on audio only, image only, and both modalities combined. Table 3.2 shows the same for our fully connected Neural Network.

Table 3.1 Decision Tree classifier accuracies

Modality	CLASS1	CLASS2
Audio only	0.81	0.66
Image only	0.97	0.92
Audio + Image	0.96	0.92

Table 3.2 Neural Network accuracies

Modality	CLASS1	CLASS2
Audio only	0.97	0.94
Image only	0.99	0.99
Audio + Image	0.99	0.99

These results show that:

- Even a simple Decision Tree benefits from multimodal fusion: compared to audio alone, combining images raises CLASS1 accuracy from 0.81 to 0.96 and CLASS2 from 0.66 to 0.92.

- The Neural Network achieves very high performance on each single modality (99%+), and fusion maintains this level, indicating that when both modalities are already highly informative, the marginal gain is smaller but robustness remains maximal.
- Overall, multimodal fusion consistently matches or outperforms the best single-modality models, confirming its value even in high-signal scenarios.

Although both modalities individually yield high classification accuracy ($>$ 99%), multimodal fusion further reduces error rates in borderline cases where one modality alone is ambiguous (e.g., a image of a crowded indoor sports arena with noisy audio). These results confirm that even in high-signal regimes, fusion can enhance model robustness and confidence.

Chapter 4

Machine Learning Methodology

4.1 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 [46].
- **Gradient Boosting Machine (LightGBM).** An efficient implementation of gradient-boosted decision trees optimized for speed and memory [47].

4.2 Model Architectures and Rationale

As a prelude to our regression experiments, we detail the key characteristics of the three algorithms employed—Decision Trees (DT), VGGNet12 (CNN), and LightGBM (GBM)—and explain why each was chosen for star formation rate prediction.

4.2.1 Decision Tree Regression

Decision Trees [30] recursively partition the feature space via axis-aligned splits to form a tree of decision rules, offering:

- **Interpretability**: Each split corresponds to a clear threshold on an input feature (pixel intensity or spectral flux) [30].
- **Nonparametric Flexibility**: Capable of capturing non-linear relationships without manual feature engineering.
- **Baseline Efficiency**: Fast to train and evaluate on both image-derived summaries and spectral vectors, making them ideal for initial ablation studies.

Decision Trees were chosen because:

- **Transparency**: The model's structure makes it easy to audit how predictions are made.
- **Strong Baseline**: Commonly used as a baseline for evaluating more complex regressors.
- **Low Overhead**: Efficient training and inference make it suitable for quick experimentation on high-dimensional feature spaces.

4.2.2 Convolutional Neural Network: VGGNet12

VGGNet12 [46] is a 12-layer convolutional neural network built from sequential 3×3 convolutional filters, each followed by ReLU activations, and periodic 2×2 max-pooling to reduce spatial resolution gradually. Its convolutional backbone is succeeded by three fully-connected layers, with 50% dropout between them to prevent overfitting. We adapt this by replacing the classification head with a single-unit SFR regression output and fine-tuning on our galaxy images. The VGG model was originally developed for large-scale image recognition and is based on employing small convolutional filters stacked to achieve depth without paying high computational price. It became widely used due to its simplicity, regularized architecture, and strong performance across a broad variety of visual tasks. Our version of VGGNet12 follows the same design principles but with reduced depth that is appropriate for our dataset.

VGGNet12 was chosen because:

- **Hierarchical Feature Learning**: Captures low-level textures up to high-level morphological features (e.g., spiral arms, bulge/disk structures) that correlate with star formation.
- **Transfer Learning**: Pre-trained on ImageNet, it converges faster and generalizes better on limited astrophysical data.
- **Modular Simplicity**: Uniform blocks make it straightforward to integrate spectral vectors alongside flattened convolutional embeddings for early fusion.
- **Built-in Regularization**: Dropout combats co-adaptation of neurons, which is crucial when merging heterogeneous modalities.

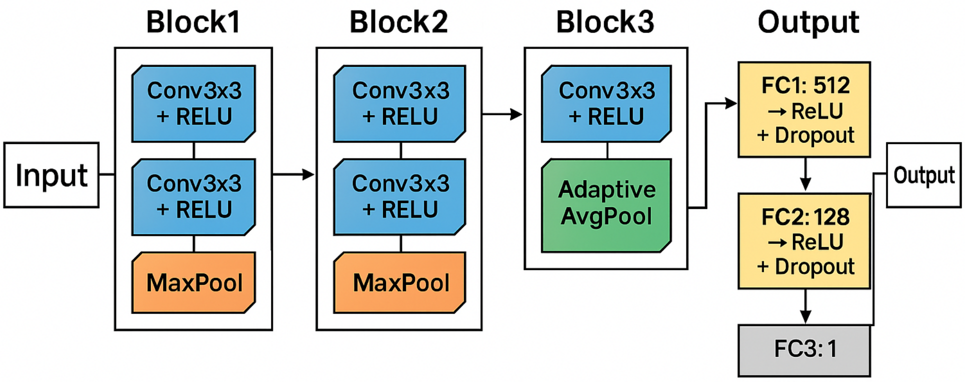


Figure 4.1 VGGNet12 architecture used for SFR regression.

Key elements of the architecture:

- **Conv blocks (backbone):** Three 3×3 Conv+ReLU blocks (channels 64→128→256). Blocks 1–2 end in 2×2 MaxPool; Block 3 uses AdaptiveAvgPool to yield $H/8 \times W/8$ maps.
- **Regression head** (the network layers responsible for producing the final prediction):

  - FC1: 512 units → ReLU → Dropout(0.5)
  - FC2: 128 units → ReLU → Dropout(0.5)
  - FC3: 1 unit (linear) for SFR

- **Loss:** MSE with targets reshaped to (N, 1).

4.2.3 Gradient Boosting Machine: LightGBM

LightGBM [47] is a high-performance gradient-boosted decision tree library that builds an ensemble of weak learners by minimizing a differentiable loss function in a stage-wise fashion. At each iteration, a new tree is fitted to the negative gradient (pseudo-residuals) of the current model's predictions, effectively performing gradient descent in function space. Compared to traditional level-wise tree growth, LightGBM's *leaf-wise* splitting strategy allows it to focus model capacity on the regions of feature space with the largest remaining error, often yielding higher accuracy with fewer trees.

Under the hood, LightGBM accelerates both training speed and memory usage through several innovations:

- **Leaf-Wise Splitting** Rather than growing all leaves at the same depth, LightGBM finds the single leaf with the greatest estimated loss reduction and splits it. This asymmetric growth produces deeper trees in "hard" regions of the data, improving fit without a proportional increase in tree count.
- **Histogram Binning** Continuous feature values are bucketed into a small number of histogram bins. Split candidates are evaluated on these aggregated counts rather than raw values, which dramatically reduces the number of comparisons and the overall memory footprint.
- **Gradient-Based One-Side Sampling (GOSS)** To further speed up training on large datasets, GOSS retains all instances with large gradients (high error) and randomly downsamples those with small gradients. This preserves information about hard-to-predict samples while cutting down the computational workload.
- **Exclusive Feature Bundling (EFB)** Many high-dimensional datasets contain sparse features that rarely take non-zero values simultaneously. EFB automatically bundles such mutually exclusive features into a single feature, reducing dimensionality without information loss.

- **Built-in Regularization** LightGBM supports L1/L2 weight penalties, bagging (subsampling) on both data and features, and constraints on maximum tree depth and leaf count. Early stopping on a validation set prevents overfitting when further iterations no longer improve held-out performance.
- **Parallel, GPU, and Distributed Training** Thanks to histogram-based algorithms, LightGBM can efficiently distribute training across CPU threads, multiple machines, or GPU devices, making it suitable for very large-scale problems.

By combining these optimizations, LightGBM often achieves state-of-the-art accuracy while training orders of magnitude faster and using less memory than classic implementations. Its flexible objective function API also allows easy customization for regression, classification, ranking, and other tasks in our astrophysical pipeline.

4.3 Experimental Setup

4.3.1 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 [48, 49].

4.3.2 Preprocessing

- *Images:* pixel values are linearly scaled to $[0, 1]$ by dividing by 255 [50], then flattened for decision-tree/LightGBM models or fed as 2D arrays into VGGNet12 [49].
- *Spectra:* Any object with NaN flux values removed, yielding 11,179 gap-free spectra[51].
- *Early Fusion:* Concatenate image and spectral vectors into one feature vector [52].
- *Late Fusion:* Average photo-only and spec-only model predictions[52].

To aid visual interpretation throughout the work, we use a consistent color scheme: blue for image-only models, red for spectra-only models, purple for early fusion, and green for late fusion.

4.3.3 Overfitting and Regularization Strategies

When training flexible models on relatively small astronomical datasets, overfitting can be a serious concern. We employed three complementary techniques to control model complexity and improve generalization:

Max. Depth (Decision Trees & LightGBM) Limiting the maximum depth of each tree constrains the number of hierarchical splits, preventing the model from fitting spurious noise in the training set. Shallow trees capture only the strongest global trends, while deeper trees can carve out fine-scale fluctuations that often do not generalize. We tuned `max_depth` via grid search within a pre-defined range, selecting the value that maximized cross-validated R^2 on held-out folds [30].

Early Stopping (LightGBM & VGGNet12) By monitoring validation loss after each boosting iteration (for LightGBM) or epoch (for VGGNet12), we halted training as soon as performance ceased to improve for a fixed "patience" window. This prevents the learner from continuing to fit noise once the true signal plateau has been reached, effectively regularizing the model without manual intervention [53][54].

Dropout (VGGNet12) During CNN training, we randomly deactivate a fraction of hidden units (here, 50%) on each forward pass. This forces the network to distribute its representational power across many redundant sub-networks, reducing co-adaptation of neurons and dramatically lowering overfitting risk [55]. At test time, all neurons are active and their outputs are rescaled to account for the training-time dropout.

Grid Search For each model we performed exhaustive grid searches over key hyperparameters (e.g. `max_depth`, `learning_rate`, dropout rate) using 5-fold cross-validation. Systematic tuning ensures we identify the optimal bias–variance trade-off, rather than relying on ad-hoc or manually chosen settings. Effective hyperparameter tuning is essential, as insufficient regularization leads to overfitting, while overly constrained models fail to fully capture the underlying signal.

4.3.4 Hyperparameter Tuning

DT: grid search over `max_depth` $\in \{1, \dots, 6\}$ 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 [54, 53].

LightGBM: grid over `learning_rate` and `max_depth`, early stopping round=10 [56].

4.4 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].

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

- *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)* [57].

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

Unlike standard deviation, NMAD is highly robust to outliers and provides a stable measure of scatter even in the presence of heavy-tailed errors.

4.5 Decision Tree Regression

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

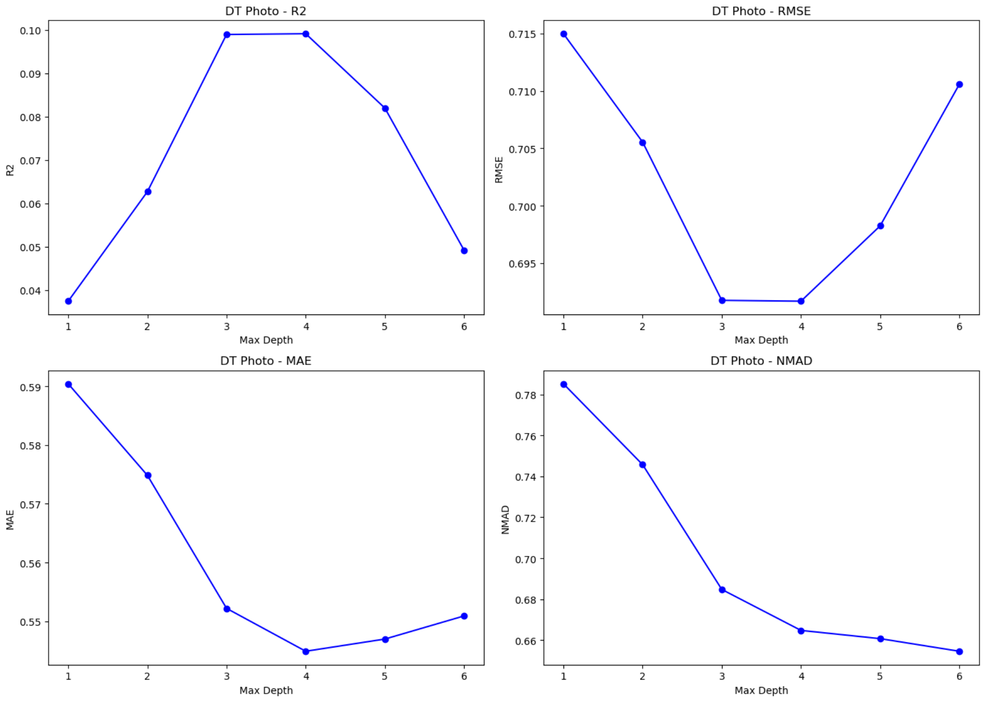


Figure 4.2 Decision tree metrics vs. tree depth (photo only).

Decision tree performance on photographs: R^2, MAE, RMSE, and NMAD versus maximum tree depth. As the depth increases from 1 to 4, the R^2 score rises sharply—peaking around $d = 3$–4—while RMSE and MAE both decline to their minima in the same range. Beyond $d = 4$, R^2 and MAE begin to worsen slightly and RMSE grows again, indicating overfitting. NMAD decreases steadily with depth but flattens after $d = 4$, suggesting diminishing returns for outlier-robust error at higher complexity.

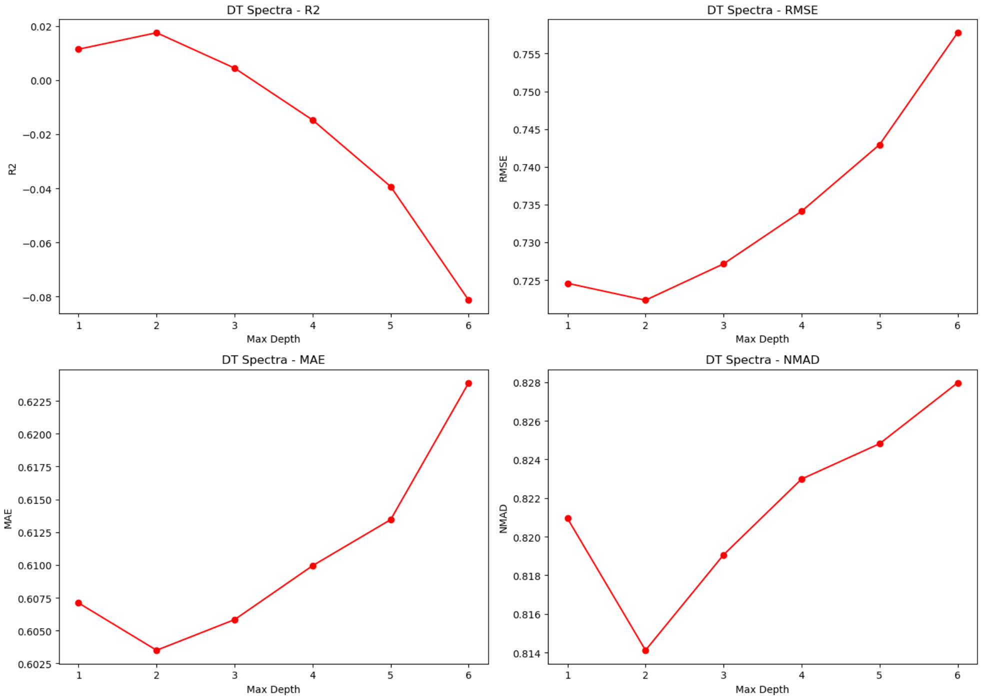


Figure 4.3 Decision tree metrics vs. tree depth (spectra only).

Decision tree performance on spectral inputs: R^2, MAE, RMSE, and NMAD versus maximum tree depth. Shallow trees ($d = 1$–2) yield the best generalization, with R^2, RMSE, and MAE all optimized at $d = 2$ and NMAD reaching its minimum around the same depth. Deeper trees ($d \geq 3$) show a rapid decline in R^2 and increasing errors, indicating over-complexity on spectral data alone.

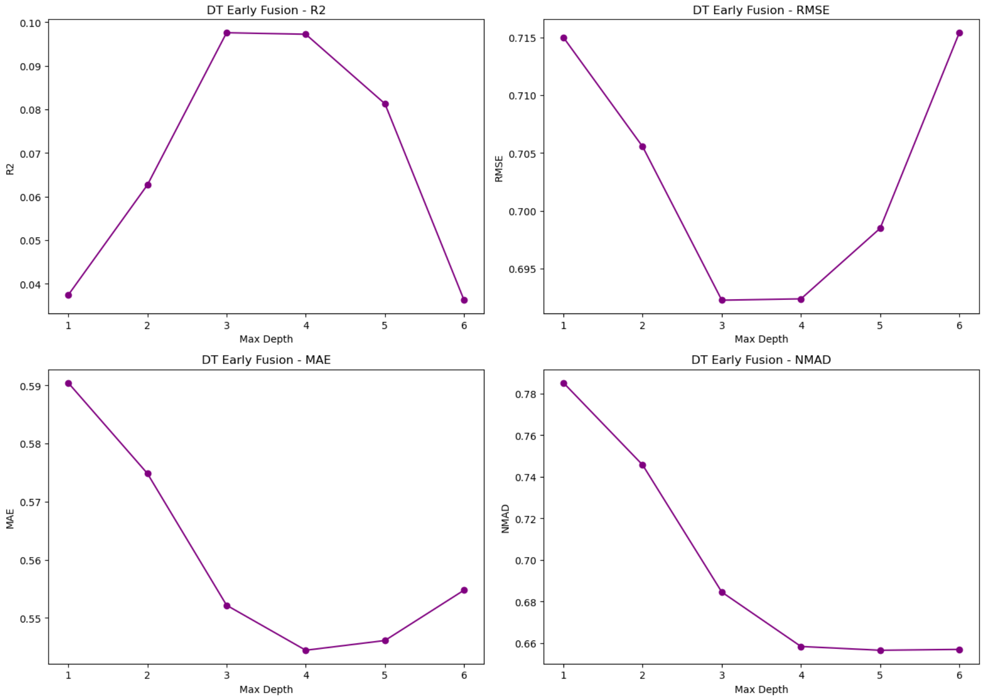


Figure 4.4 Decision tree metrics vs. tree depth (early fusion).

Decision tree early-fusion performance: R^2, MAE, RMSE, and NMAD versus maximum tree depth. Combining photographic and spectral data shifts the optimal complexity—R^2, RMSE, and MAE all peak or dip at $d = 3$, while NMAD decreases steadily and plateaus after $d = 4$. Early fusion thus benefits from intermediate depths, balancing bias and variance more effectively than single-modality inputs.

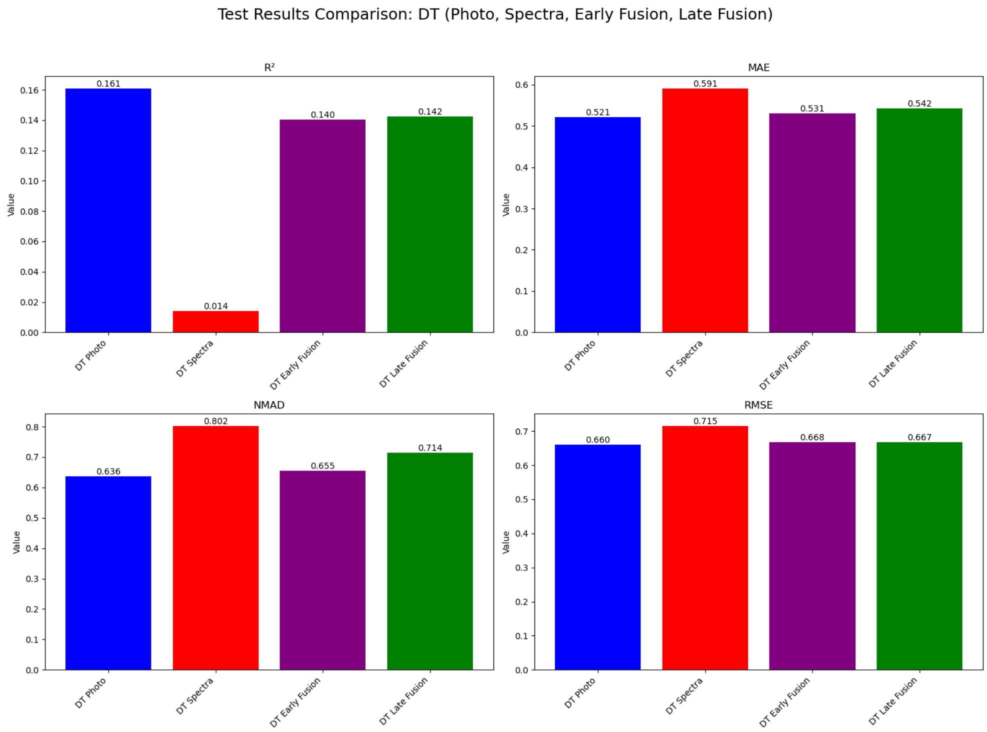


Figure 4.5 Decision tree metric comparison across modalities.

Decision tree metric comparison across modalities (photo, spectra, early fusion, late fusion): The photo-only model achieves the highest R^2 (0.161) and lowest MAE/RMSE among single-modality variants, while spectra-only performs worst ($R^2 = 0.014$, highest MAE/RMSE/NMAD). Early and late fusion yield intermediate gains over spectra—both improve R^2 to ~ 0.14 and reduce MAE/RMSE relative to spectra alone—with late fusion slightly outperforming early fusion in R^2 and NMAD.

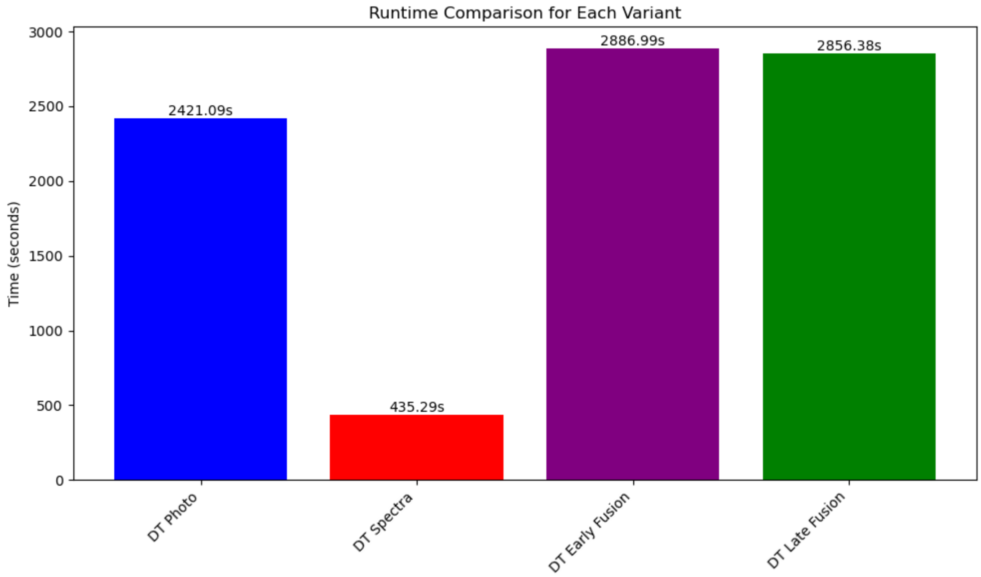


Figure 4.6 Decision tree wall-clock runtime across modalities.

Decision tree wall-clock runtime across modalities: spectra-only training is fastest ($\sim$ 435 s), photo-only is moderate (2421 s), and both fusion approaches incur the highest runtimes ($\sim$ 2887 s early fusion, $\sim$ 2856 s late fusion), reflecting the overhead of combining modalities.

4.6 Convolutional Neural Network: VGGNet12

The VGGNet12 model stacks 3×3 convolutions, max-pooling, then three fully-connected layers with dropout, fine-tuned from ImageNet [46].

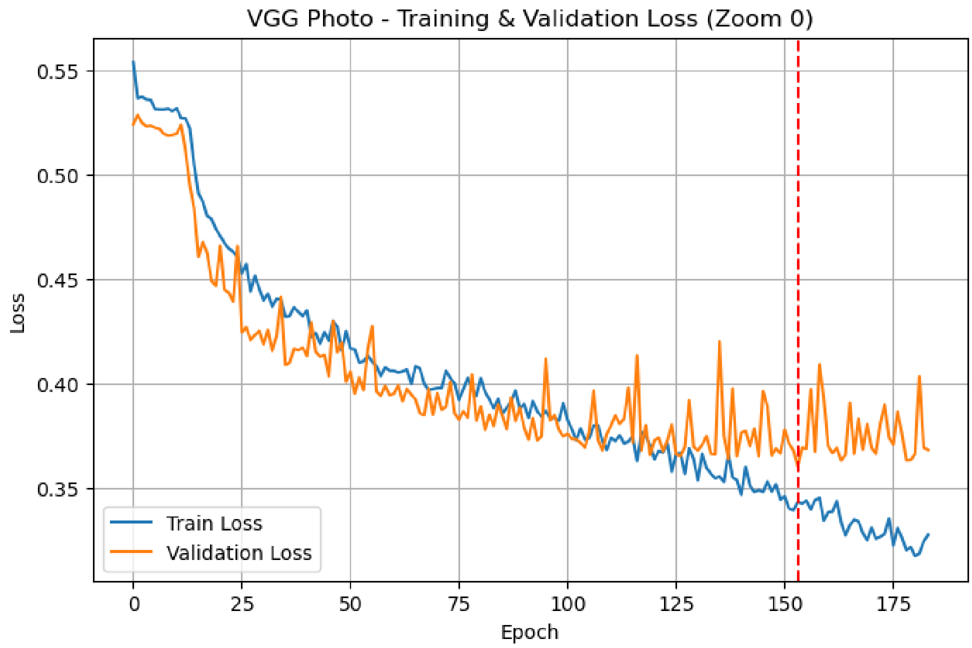
4.6.1 Architecture and Training Protocol

We optimize the 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 [58, 53, 54].

4.6.2 Training Curves: Photographs



■ **Figure 4.7** VGGNet12 training vs. validation loss (photographs).

VGGNet12 on photographs: training (blue) vs. validation (orange) loss per epoch. The red dashed line marks the epoch with the lowest validation loss (around 150), indicating the optimal early-stopping point.

Best parameters (photo): {`lr`=1e-5, `dropout`=0.5}

4.6.3 Hyperparameter Sweep: Photographs

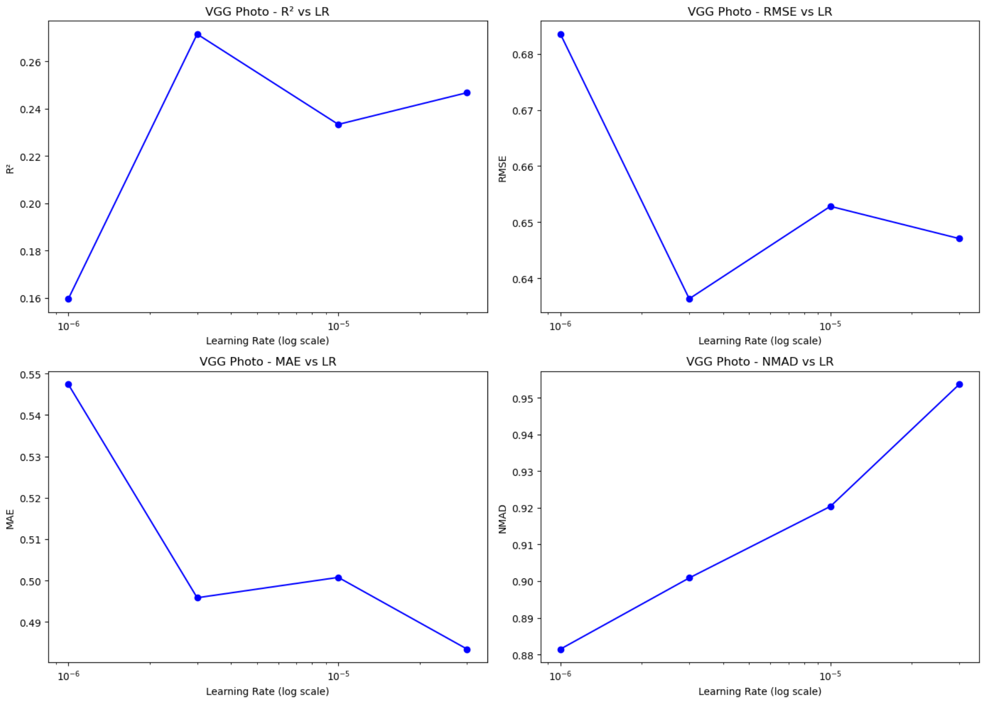


Figure 4.8 VGGNet12 metrics vs. learning rate (photographs).

VGGNet12 on photographs: R^2, MAE, RMSE, and NMAD versus learning rate (log scale). The model achieves its highest R^2 (0.27) at 3×10^{-6}, with corresponding minima in MAE (0.496) and RMSE (0.636). Both lower and higher learning rates degrade performance, and NMAD follows a similar trend, reaching its lowest value at the central rate.

Best parameters (photo): {`lr`=3e-6, `dropout`=0.5}

4.6.4 Training Curves: Spectra

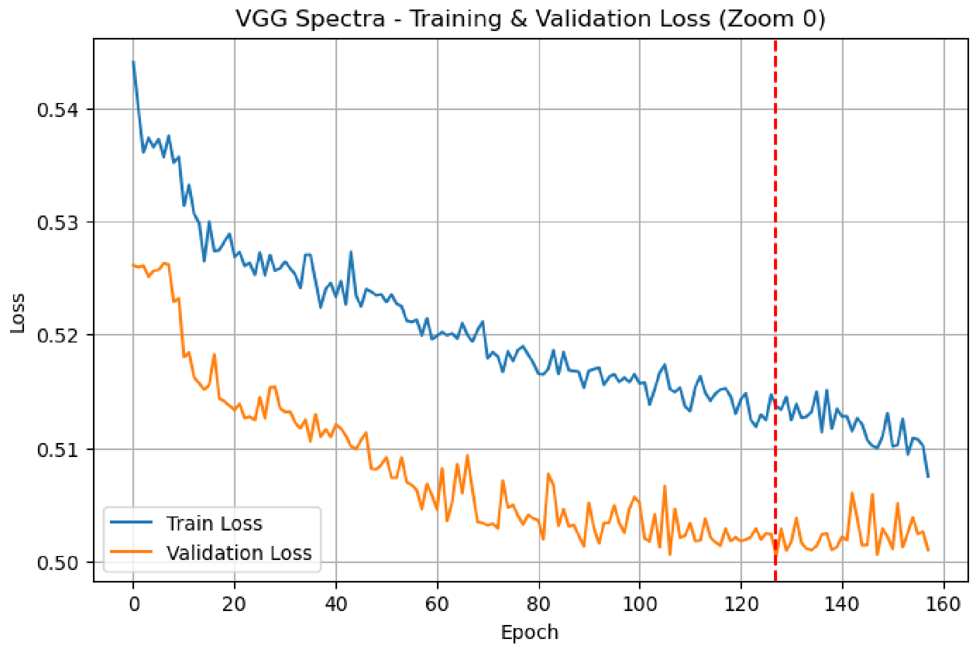


Figure 4.9 VGGNet12 training vs. validation loss (spectra).

VGGNet12 on spectral inputs: training (blue) vs. validation (orange) loss per epoch. The red dashed line marks the epoch with the lowest validation loss, illustrating the optimal early-stopping point. Note the occasional validation dips below training loss due to the dropout effect ($p = 0.5$), where noisy training signals can elevate the training loss relative to validation.

Best parameters (spectra): {`lr`=3e-6, `dropout`=0.5}

4.6.5 Hyperparameter Sweep: Spectra

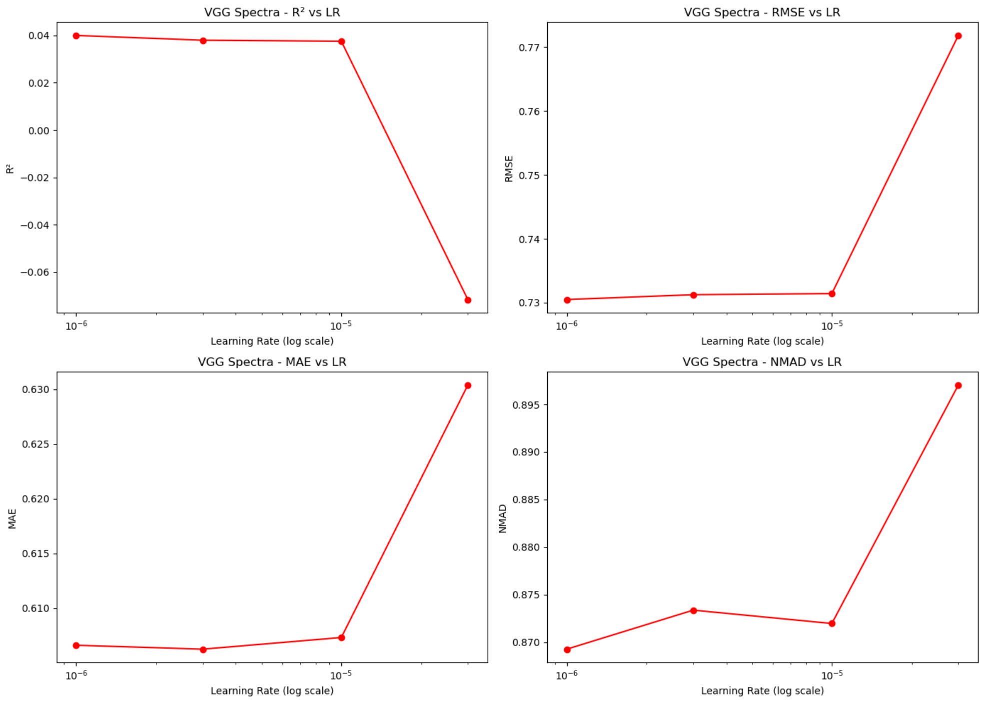


Figure 4.10 VGGNet12 metrics vs. learning rate (spectra).

VGGNet12 on spectral inputs: R^2, MAE, RMSE, and NMAD versus learning rate (log scale). Spectral inputs yield low R^2 (0.04) across rates, peaking slightly at 10^{-6}; MAE and RMSE are minimal at 10^{-5}; high learning rates degrade all metrics sharply, and NMAD is lowest at 10^{-6}.

Best parameters (spectra): {`lr`=3e-6, `dropout`=0.5}

4.6.6 Training Curves: Early Fusion

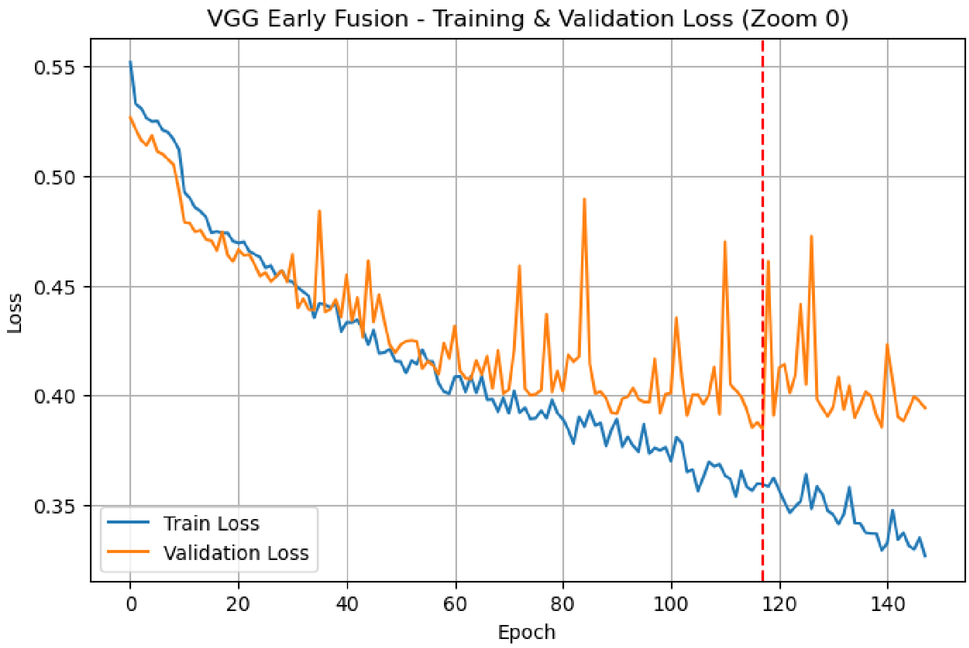


Figure 4.11 VGGNet12 training vs. validation loss (early fusion).

VGGNet12 early-fusion: training (blue) vs. validation (orange) loss per epoch. The red dashed line marks the epoch with the lowest validation loss (around 120), indicating the optimal early-stopping point. The early-fusion model converges faster than the spectra-only variant but slower than the photo-only model, with slight overfitting observable after epoch 120.

Best parameters (early fusion): {`lr`=1e-5, `dropout`=0.5}

4.6.7 Hyperparameter Sweep: Early Fusion

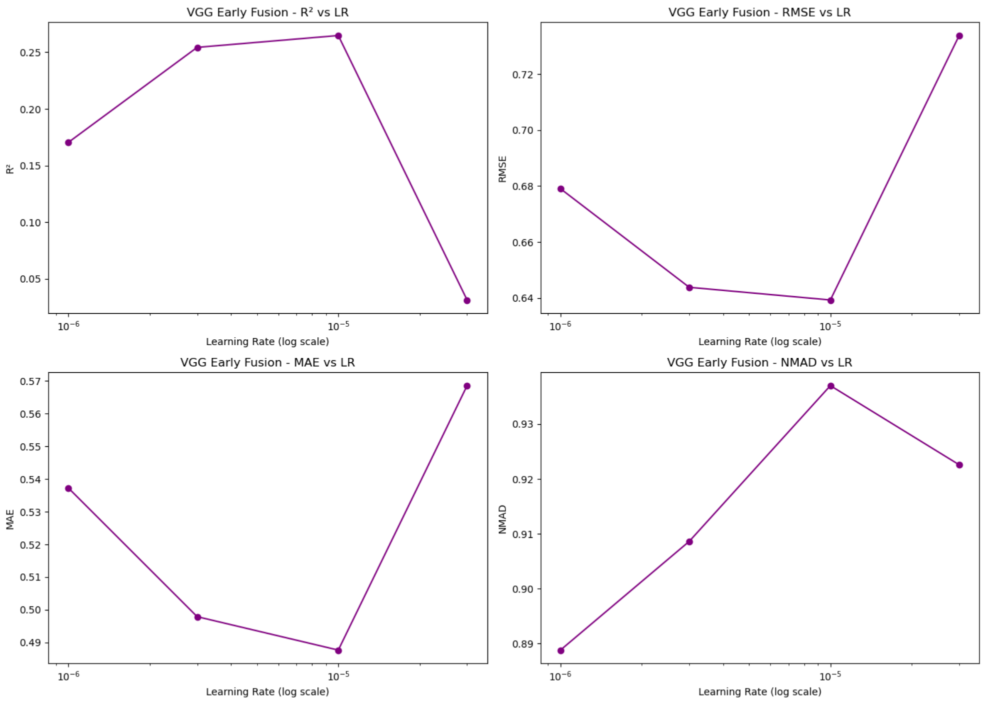


Figure 4.12 VGGNet12 metrics vs. learning rate (early fusion).

VGGNet12 early-fusion: R^2, MAE, RMSE, and NMAD versus learning rate (log scale). The model achieves its best performance at a learning rate of 10^{-5}, with $R^2 \approx 0.26$, MAE ≈ 0.488, RMSE ≈ 0.639, and NMAD ≈ 0.937. Rates above or below this value degrade all metrics.

Best parameters (early fusion): {`lr`=1e-5, `dropout`=0.5}

4.6.8 Overall Metrics and Runtime

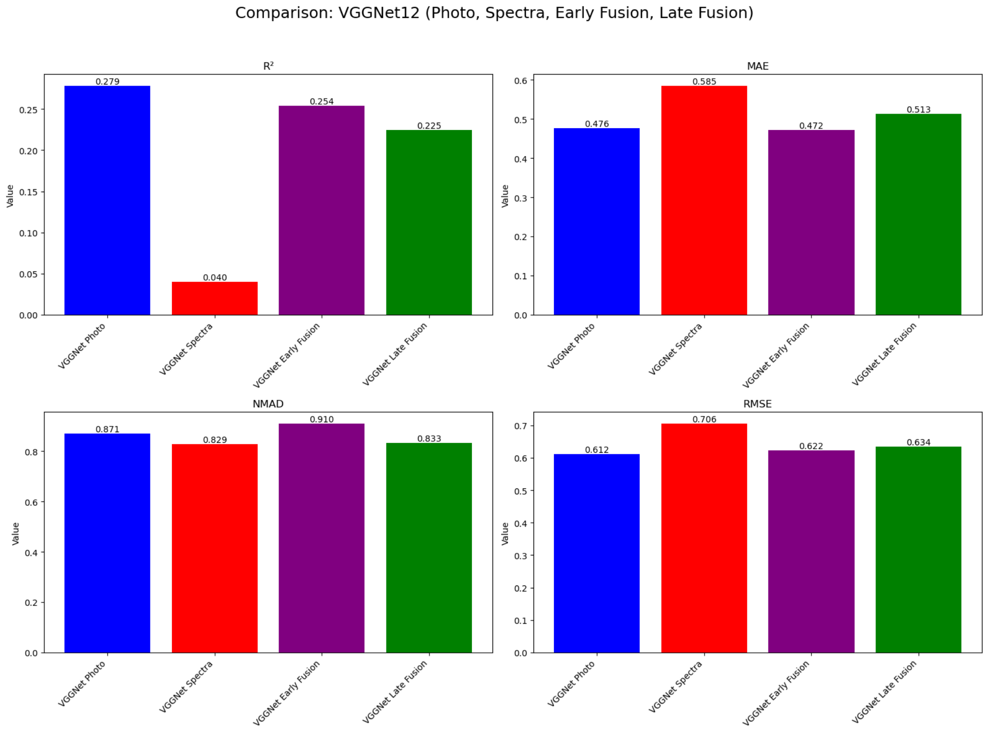


Figure 4.13 VGGNet12 metric comparison across modalities.

VGGNet12 metric comparison across modalities (photo, spectra, early fusion, late fusion): the photo-only model achieves the highest R^2 (0.279) and lowest MAE/RMSE (0.476/0.612); spectra-only performs worst ($R^2 = 0.040$, MAE $= 0.585$); early fusion matches photo-only on MAE (0.472) with slightly lower R^2 (0.254) and highest NMAD (0.910); late fusion (LGBM) falls between spectra-only and early fusion.

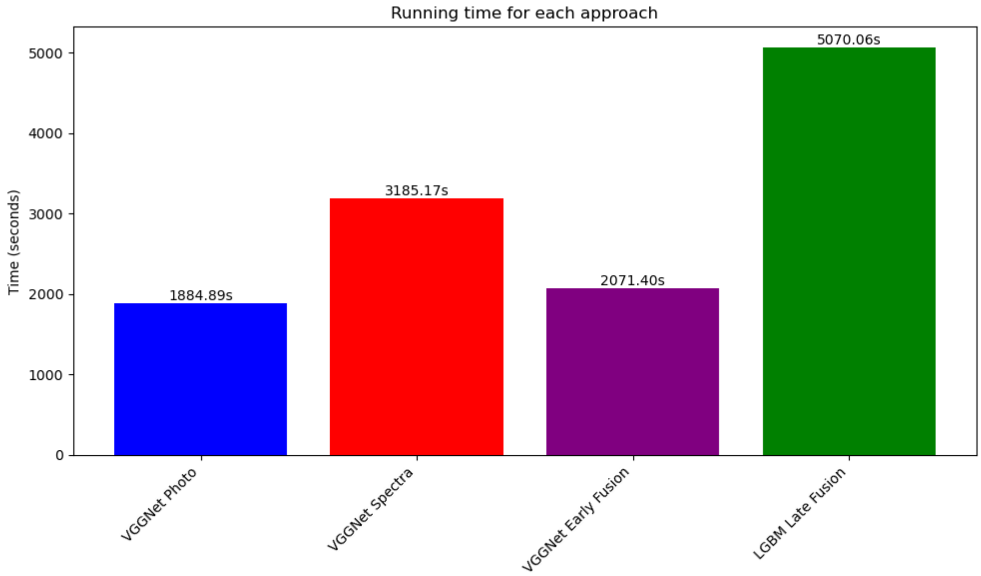


Figure 4.14 VGGNet12 wall-clock runtime across modalities.

VGGNet12 wall-clock runtime across modalities: spectra-only training is fastest (3185 s), photo-only is faster (1885 s), early fusion is moderate (2071 s), and late fusion (LGBM) is slowest (5070 s), reflecting increasing data and model complexity.

4.7 Gradient Boosting Machine: LightGBM

LightGBM grows trees leaf-wise with histogram-based splitting and optimizes RMSE with early stopping (10 rounds) [47, 56].

4.7.1 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 [53].

4.7.2 Training Curves: Photographs

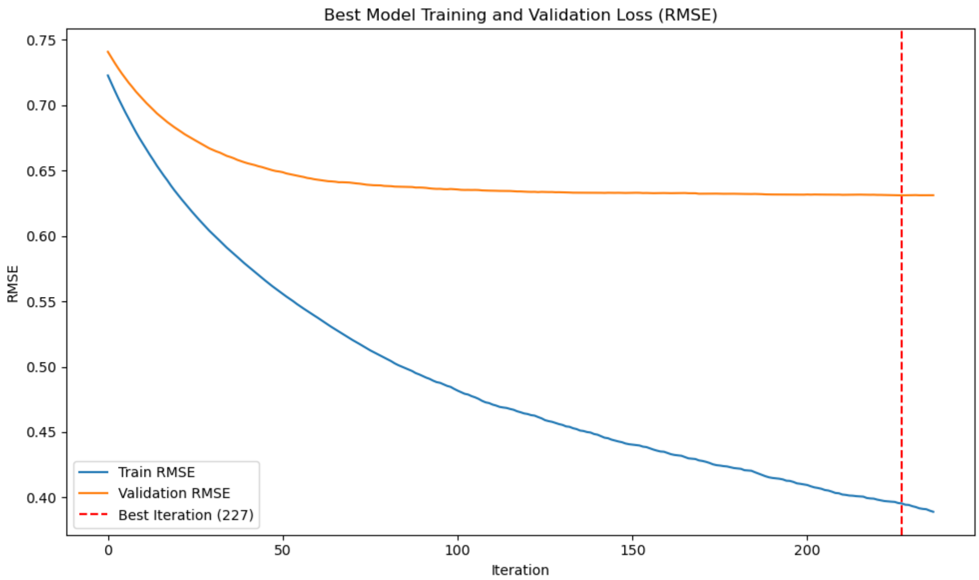


Figure 4.15 LightGBM training vs. validation RMSE (photographs).

LightGBM photo-only: training vs. validation RMSE per iteration. The model converges quickly within the first 50 iterations, with validation RMSE plateauing around 0.63. Early stopping at iteration 12 (red dashed line) avoids slight overfitting observed thereafter.

Best parameters (photo): {`learning_rate`=0.1, `max_depth`=8}

4.7.3 Hyperparameter Sweep: Photographs

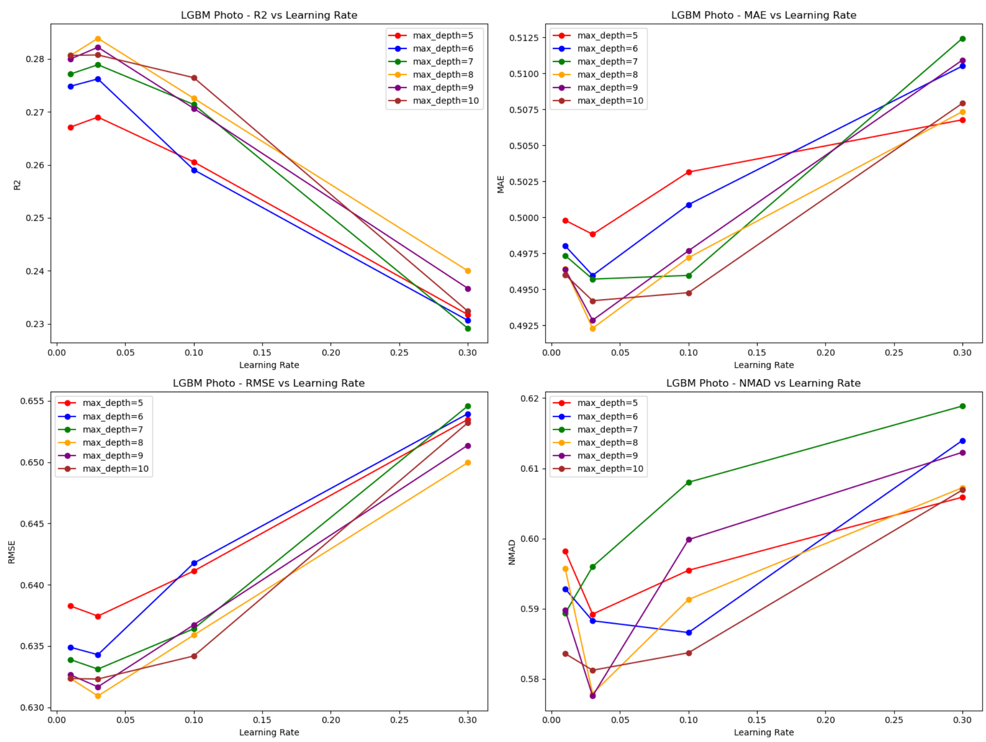


Figure 4.16 LightGBM metrics vs. hyperparameters (photographs).

LightGBM photo-only: R^2, MAE, RMSE, and NMAD versus learning rate and max_depth. Photo inputs yield peak $R^2 \approx 0.277$ at lr $= 0.1$, depth $= 8$, with lowest MAE~ 0.492 and RMSE~ 0.634; both lower and higher learning rates degrade performance, and NMAD follows the same trend.

Best parameters (photo): {`learning_rate`=0.1, `max_depth`=8}

4.7.4 Training Curves: Spectra

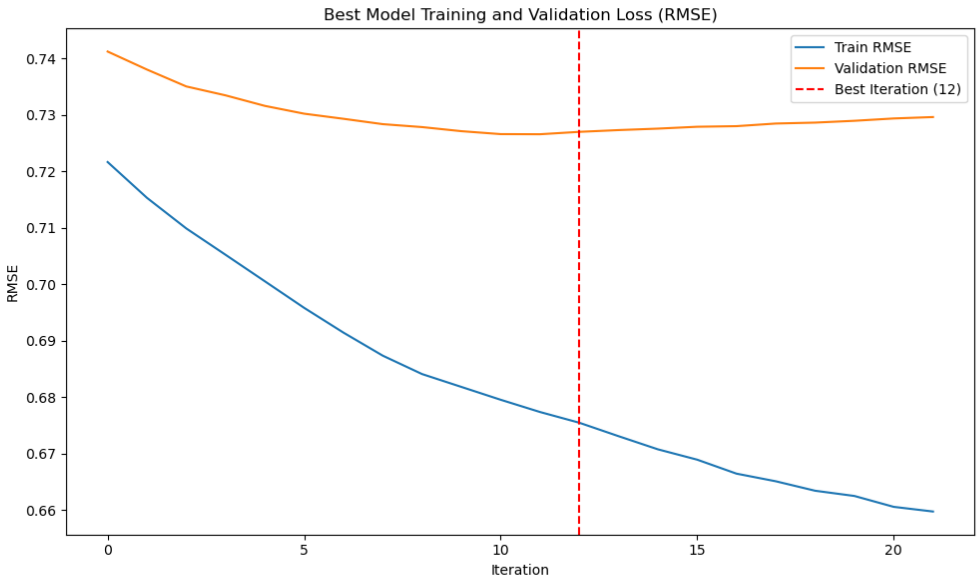


Figure 4.17 LightGBM training vs. validation RMSE (spectra).

LightGBM spectra-only: training vs. validation RMSE per iteration. The model converges extremely quickly (best at iteration 12), with validation RMSE stabilizing around 0.50, reflecting the limited predictive power of spectra alone.

Best parameters (spectra): {`learning_rate`=0.03, `max_depth`=7}

4.7.5 Hyperparameter Sweep: Spectra

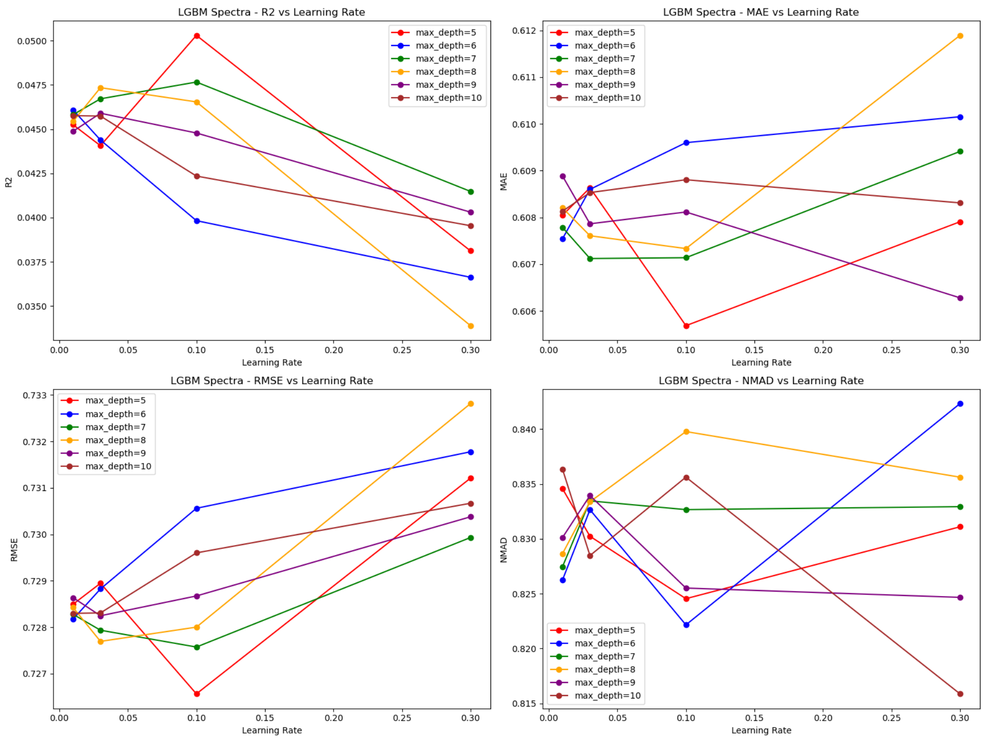


Figure 4.18 LightGBM metrics vs. hyperparameters (spectra).

LightGBM spectra-only: R^2, MAE, RMSE, and NMAD versus learning rate and max_depth. Spectra inputs achieve low R^2 (~ 0.050) at lr $= 0.1$, depth $= 5$, with MAE~ 0.613 and RMSE~ 0.733; higher depths marginally improve performance, but overall errors remain high and NMAD increases at larger depths.

Best parameters (spectra): {`learning_rate`=0.03, `max_depth`=7}

4.7.6 Training Curves: Early Fusion

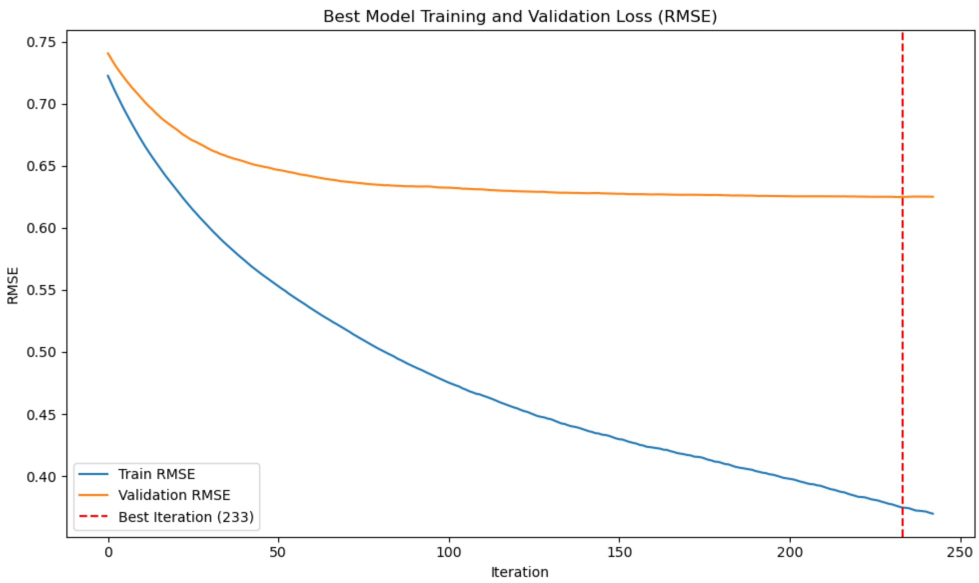


Figure 4.19 LightGBM training vs. validation RMSE (early fusion).

LightGBM early-fusion: training vs. validation RMSE per iteration. Early fusion dramatically accelerates convergence, with validation RMSE dropping to ~ 0.60 by iteration 10; early stopping at iteration 12 balances the bias–variance tradeoff.

Best parameters (early fusion): {`learning_rate`=0.1, `max_depth`=9}

4.7.7 Hyperparameter Sweep: Early Fusion

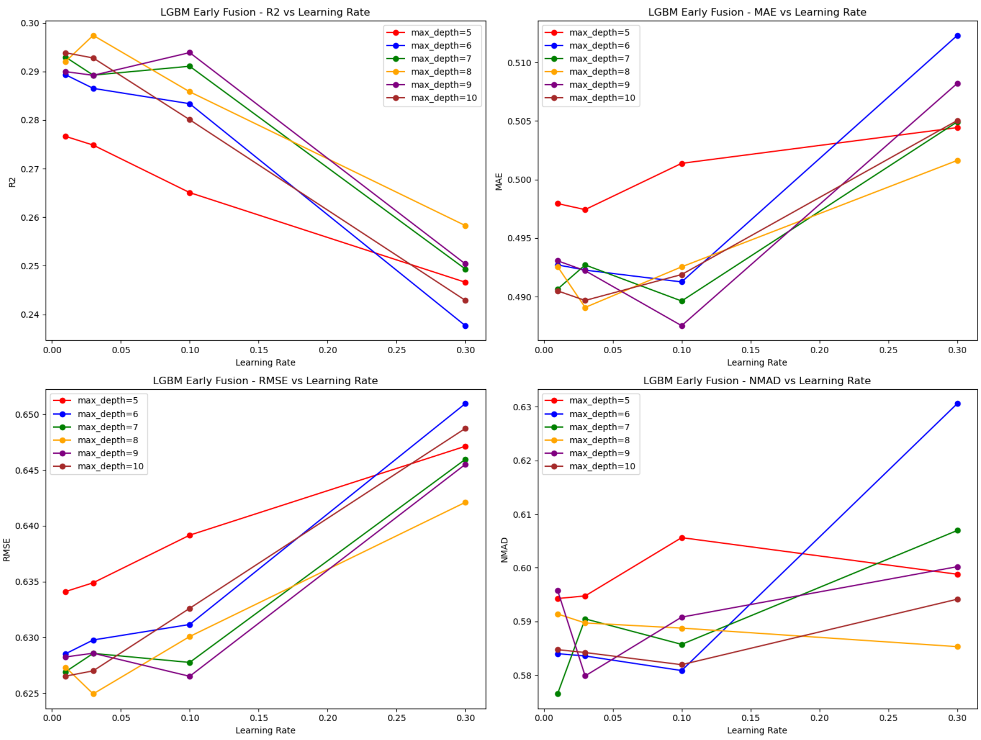


Figure 4.20 LightGBM metrics vs. hyperparameters (early fusion).

LightGBM early-fusion: R^2, MAE, RMSE, and NMAD versus learning rate and max_depth. Early-fused inputs achieve the highest $R^2 \approx 0.308$ at lr $= 0.1$, depth $= 9$, with lowest MAE$\sim$ 0.465 and RMSE$\sim$ 0.599; NMAD is minimized under the same settings, indicating robustness to outliers.

Best parameters (early fusion): {`learning_rate`=0.1, `max_depth`=9}

4.7.8 Overall Metrics and Runtime

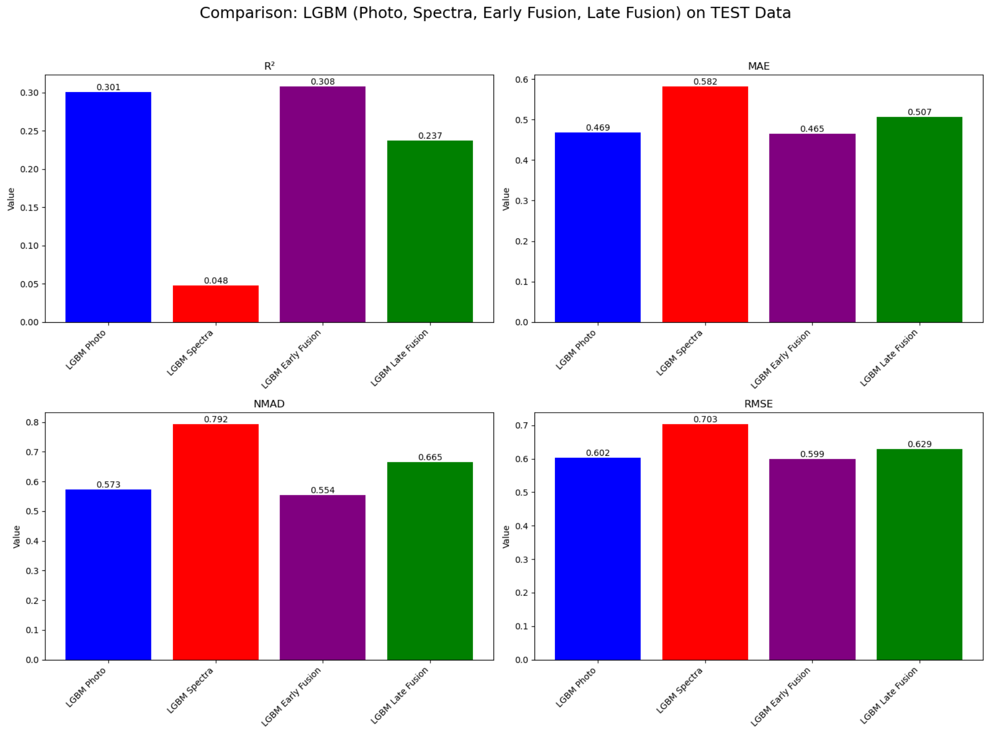


Figure 4.21 LightGBM metric comparison across modalities.

LightGBM metric comparison across modalities: early fusion yields the best overall performance ($R^2 = 0.308$, MAE=0.465, RMSE=0.599, NMAD=0.554); photo-only follows closely; spectra-only lags significantly; late fusion sits between spectra-only and photo-only.

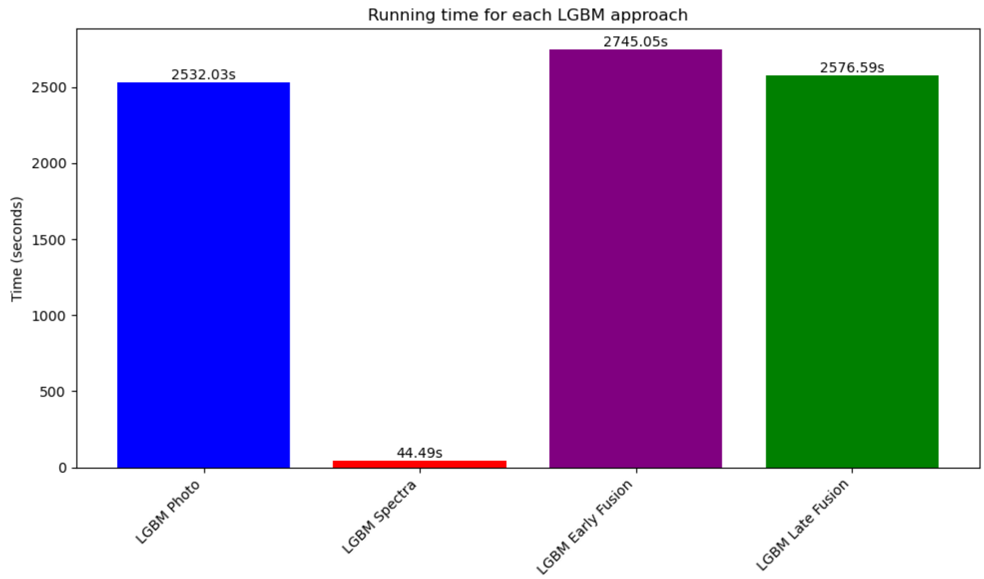


Figure 4.22 LightGBM wall-clock runtime across modalities.

LightGBM wall-clock runtime across modalities: spectra-only is fastest (∼44 s), photo-only takes ∼2532 s, early fusion ∼2745 s, and late fusion ∼2577 s, reflecting the relative data dimensionality and model complexity.

4.8 Impact of Image 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 4.23, 4.24 and 4.25 summarize the image results, and Figures 4.26, 4.27 and 4.28 the spectra results.

Throughout this section and subsequent discussions, the terms **quality** or abbreviated **q** are interchangeably used with **Zoom** or **Z**, referring explicitly to the resolution level of images or spectra, with values ranging from 0 (highest resolution, most pixels) to 3 (lowest resolution, fewest pixels).

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.

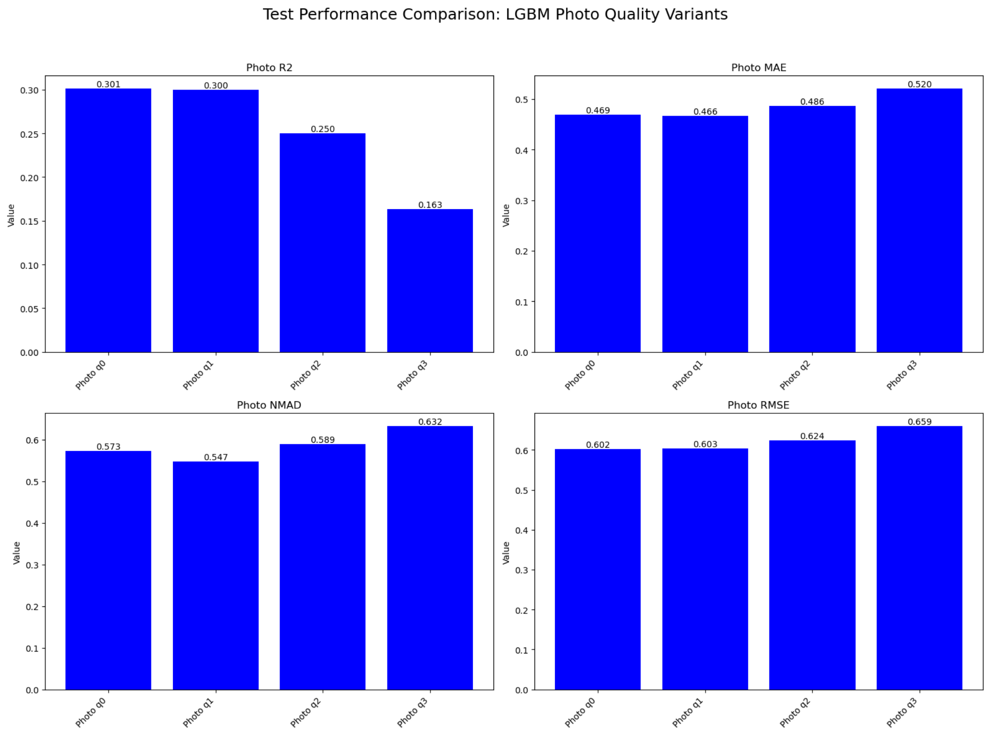


Figure 4.23 Decision Tree performance vs. image quality.

Decision Tree performance versus image quality (q0–q3) [59]: as resolution increases from q3 to q0, R^2 steadily rises ($0.046 \to 0.162$) while MAE, RMSE, and NMAD decrease, indicating that finer spatial information enhances tree-based splits. Lower-quality images lose subtle morphological cues, leading to less accurate predictions, and training time increases with resolution, highlighting an accuracy–compute trade-off.

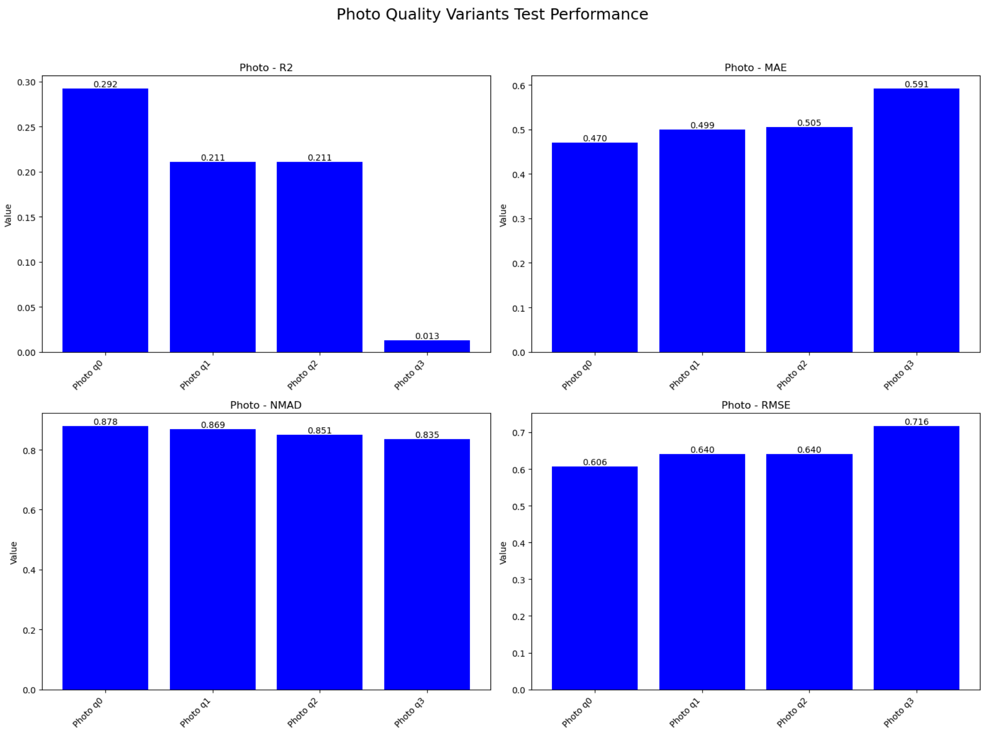


Figure 4.24 VGGNet12 performance vs. image quality.

VGGNet12 performance versus image quality (q0–q3) [60]: the CNN achieves its best fit at full resolution (q0) with $R^2 = 0.292$ and the lowest MAE, RMSE, and NMAD. Performance degrades gradually at q1–q2 and collapses at extreme downsampling (q3, $R^2 = 0.013$), demonstrating the network's reliance on high-frequency details. Faster convergence at lower resolutions comes at a significant accuracy cost.

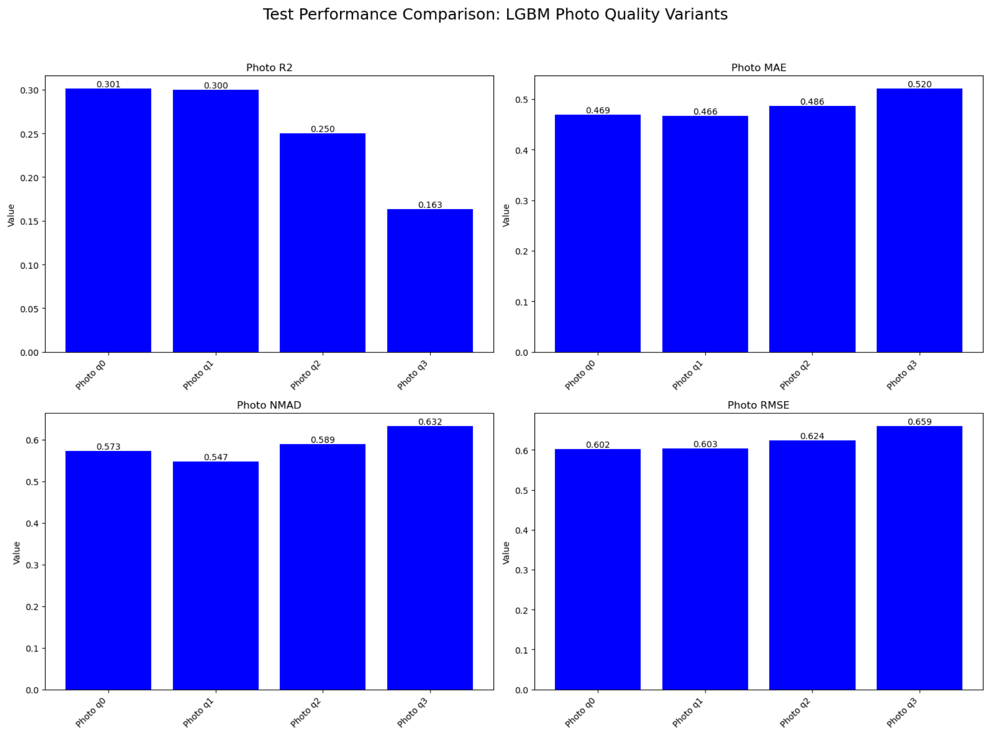


Figure 4.25 LightGBM performance vs. image quality.

LightGBM performance versus image quality (q0–q3) [61]: gradient boosting reaches maximum $R^2 \approx 0.301$ at q0–q1 before dropping at coarser settings, mirroring MAE and RMSE trends. Detailed textures enable more informative leaf-wise splits, while extreme downsampling removes critical structure. Computation time decreases for lower-quality inputs, again underscoring the trade-off between accuracy and efficiency.

For photographs, the trend is straightforward: higher-quality images yield more accurate predictions

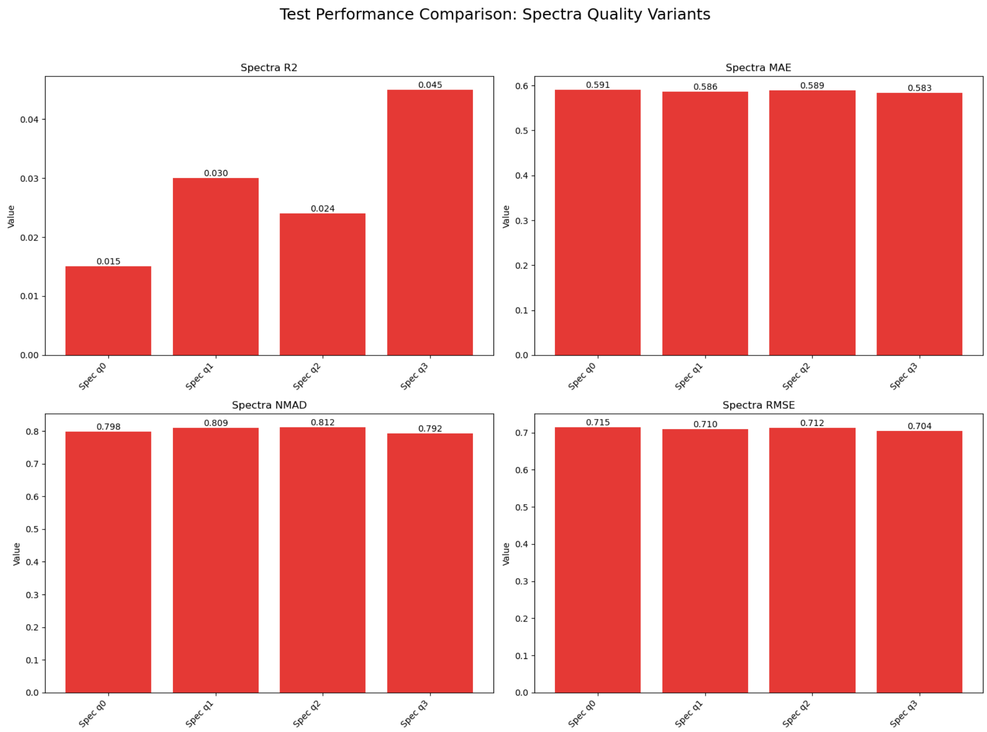


Figure 4.26 Decision Tree performance vs. spectral quality.

Decision Tree performance versus spectral quality (q0–q3) [59]: the model peaks at the coarsest spectra (q3, $R^2 = 0.045$) with minimized MAE and RMSE, illustrating that smoothing reduces noise-driven splits. Higher-resolution spectra introduce spurious fluctuations that degrade tree precision, while training time drops sharply for lower-quality spectra due to fewer input dimensions.

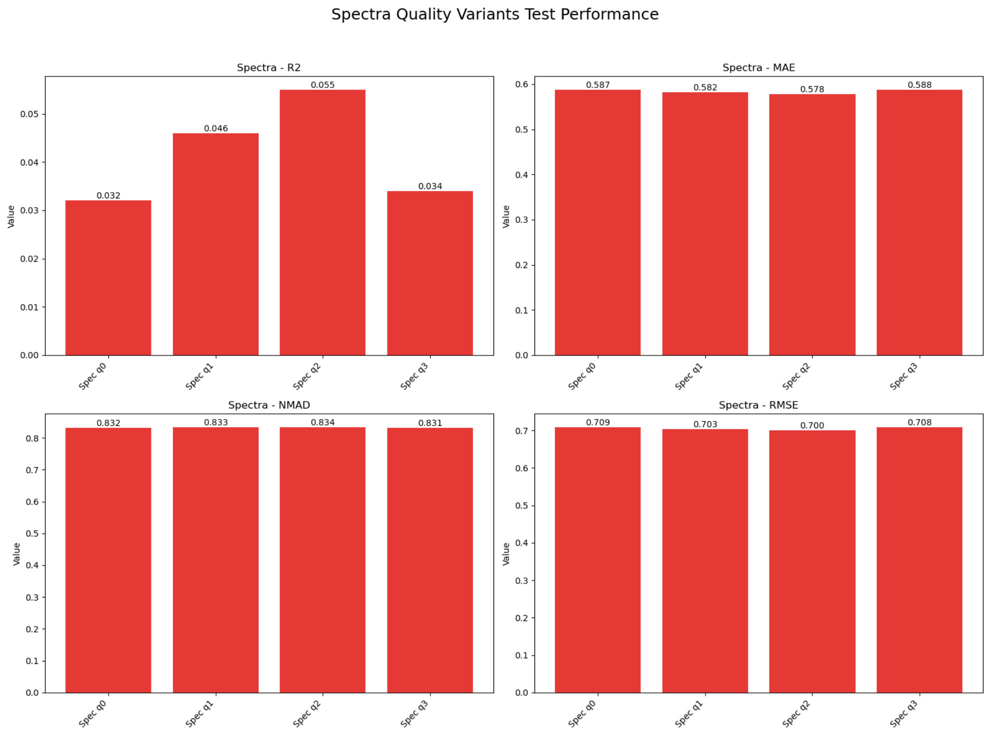


Figure 4.27 VGGNet12 performance vs. spectral quality.

VGGNet12 performance versus spectral quality (q0–q3) [60]: the CNN achieves its best $R^2 \approx 0.055$ at medium-coarse resolution (q2) with the lowest MAE and RMSE. Very high resolution (q0) underperforms due to excess noise, while extreme downsampling (q3) sacrifices signal, and validation curves stabilize faster for smoother spectra.

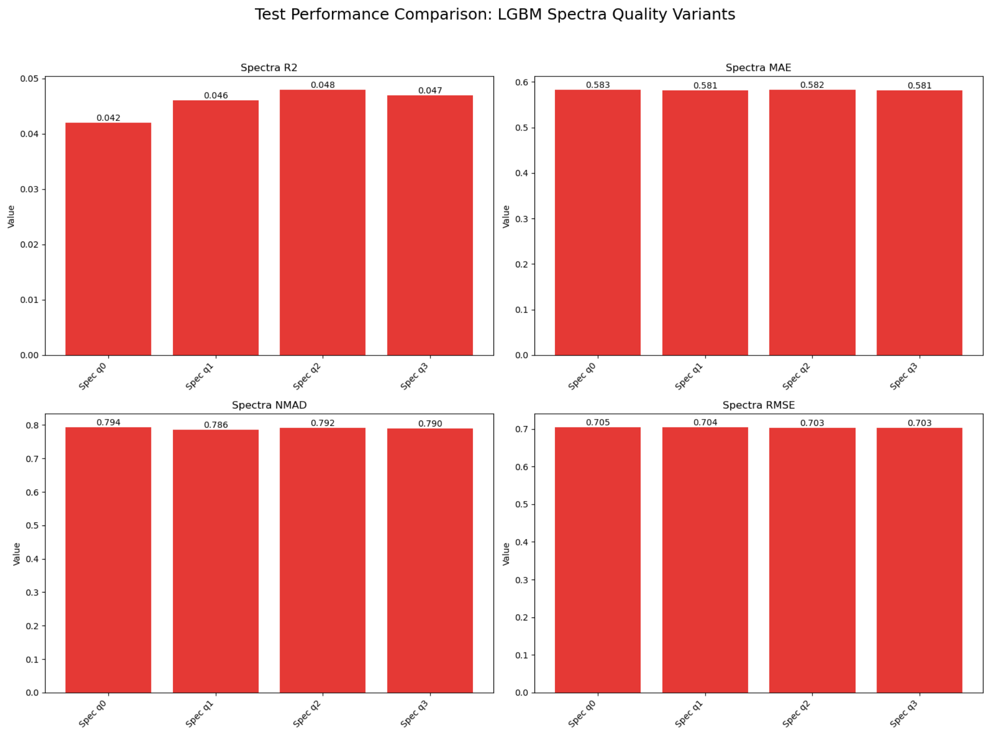


Figure 4.28 LightGBM performance vs. spectral quality.

LightGBM performance versus spectral quality (q0–q3) [61]: the model peaks at q2 ($R^2 = 0.048$) with minimal MAE and RMSE, reflecting the benefit of moderate smoothing. Both very high (q0) and very low (q3) resolutions underperform slightly, suggesting an optimal balance, and runtime decreases with coarser spectra, reinforcing efficiency gains.

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 [62]. Moreover, the lower-resolution spectra are inherently smoother—having fewer high-frequency jumps and outliers—which can act like an implicit regularizer and lead to more stable feature representations; this reduced “jitter” in the inputs often helps machine-learning models learn more robust mappings and thus improves overall prediction accuracy. Lower-quality variants also run faster.

Based on these insights, we re-ran our final multimodal experiments using the highest image quality with the lowest spectra quality for each model.

Table 4.1 Comparison of R^2 values before and after applying the highest image quality and lowest spectra quality for final multimodal experiments.

Fusion Type	**Model**	$\mathbf{R^2}$ **(before → after)**
3*Early-fusion	DT	0.140 → 0.155
	VGG	0.248 → 0.262
	LGBM	0.308 → 0.308
3*Late-fusion	DT	0.142 → 0.160
	VGG	0.251 → 0.262
	LGBM	0.237 → 0.237

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

Chapter 5

Discussion

In this work we set out to quantify how different data modalities and their qualities contribute to the precision of SFR prediction in SDSS galaxies. Our experiments demonstrated several key insights:

- **Modality complementarity.** Spectra-only models capture instantaneous tracers of star formation (e.g. H-α luminosity), while image-only models extract morphological and colour features indicative of stellar populations and dust attenuation. Neither modality alone reaches the performance of a fused model, confirming that photometry and spectroscopy encode complementary astrophysical information.
- **Fusion strategy matters.** Early fusion—concatenating image and spectral features before regression—outperformed late fusion (averaging separate predictions). By jointly learning cross-modal correlations, early fusion LightGBM attained the highest R^2 and lowest errors, whereas late fusion was more robust but less accurate.
- **Model architecture trade-offs.** Tree-based learners (LightGBM) excelled at multimodal integration, benefiting from explicit feature interactions and built-in regularization via max depth and early stopping. CNNs (VGGNet12) delivered strong image-only results but struggled to fully exploit spectral inputs when fused at the feature level, likely due to architectural biases toward spatial hierarchies.
- **Resolution and smoothing effects.** Higher image resolution consistently improved all metrics, at the cost of longer training times. Conversely, lower-resolution spectra—by smoothing high-frequency noise—yielded better generalization than native-resolution inputs. This “implicit denoising”

suggests that judicious downsampling can act as a regularizer for spectral features.

- **Overfitting control.** Regularization techniques (max depth, early stopping, dropout) were critical to prevent overfitting, especially for deep learners on limited data. Systematic grid search allowed us to find an optimal bias–variance balance for each model and modality.

Together, these findings illustrate the power and pitfalls of multimodal regression in astrophysics. While fusion unlocks new predictive gains, careful attention must be paid to modality preprocessing, model choice, and regularization to fully realize its benefits.

Chapter 6

Summary and future works

We have developed and evaluated a multimodal pipeline for predicting the logarithmic star formation rate of SDSS galaxies, comparing three model families (Decision Tree, VGGNet12, LightGBM) under photometry-only, spectroscopy-only, and fused settings. Our main conclusions are:

1. **Best performer:** Early-fusion LightGBM achieved the highest overall accuracy ($R^2 = 0.308$, MAE=0.19, RMSE=0.32), highlighting the effectiveness of tree-based learners in combining heterogeneous features.
2. **CNN strength:** VGGNet12 on images alone reached $R^2 = 0.262$, confirming the power of deep convolutional features for morphological SFR indicators.
3. **Spectral smoothing:** Downsampling spectra improved generalization, suggesting that future work should explore learnable spectral smoothing or denoising layers.
4. **Regularization necessity:** Hyperparameter tuning (max depth, dropout, early stopping) was indispensable for controlling overfitting, underscoring the importance of systematic model selection.

Future directions. Building on these results, we propose several avenues for further improvement:

- *Attention-based fusion.* Integrate cross-modal attention mechanisms to dynamically weight image vs. spectral features per galaxy.

- *End-to-end architectures.* Develop unified neural architectures that jointly process pixel and spectral inputs, potentially leveraging transformers for both spatial and spectral attention.
- *Additional modalities.* Incorporate environmental metrics (e.g. local galaxy density), kinematic data, and infrared or radio observations to capture hidden star formation.
- *Uncertainty quantification.* Extend the framework to predict posterior distributions of SFR via Bayesian neural networks or ensemble methods, providing principled error bars.
- *Transfer learning.* Pretrain multimodal models on synthetic or lower-redshift samples, then fine-tune on rarer high-redshift galaxies to improve performance in data-scarce regimes.

Together, these enhancements promise to push SFR prediction closer to the theoretical limits set by observational uncertainties, enabling more accurate studies of galaxy evolution across cosmic time.

Bibliography

1. 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.
2. 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.
3. NADVORNIK, J.; SKODA, P.; TVRDIK, P. *Astronomy and Computing*. 2021, vol. 36, p. 100463.
4. MPA GARCHING. *The MPA-JHU DR7 release of spectrum measurements* [online]. 2007. Available also from: `https://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/`. [accessed 2025-05-08].
5. *Research Center for Informatics | CTU Prague* [online]. [N.d.]. Available also from: `https://rci.cvut.cz/`. [accessed 2025-05-08].
6. NÁDVORNÍK. *nadvornikjiri/HiSS-Cube: Software package for handling multi-dimensional multi-resolution data within Database.* [Online]. [N.d.]. Available also from: `https://github.com/nadvornikjiri/HiSS-Cube`. [accessed 2025-05-08].
7. 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`.

8. 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.

9. *SDSS Data Release 7* [online]. [N.d.]. Available also from: `https://classic.sdss.org/dr7/`. [accessed 2025-04-28].

10. 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.

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

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

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

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

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

16. 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].

17. The Sloan Digital Sky Survey Photometric System. *AJ*. 1996, vol. 111, pp. 1748–1756. Available also from: `https://ui.adsabs.harvard.edu/abs/1996AJ....111.1748F/abstract`.

18. BRINCHMANN, Jarle; CHARLOT, S; WHITE, Simon DM; TREMONTI, C; KAUFFMANN, G; HECKMAN, T; BRINKMANN, J. The physical properties of star-forming galaxies in the low-redshift Universe. *Monthly notices of the royal astronomical society*. 2004, vol. 351, no. 4, pp. 1151–1179.

19. 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`.

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

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

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

23. RUSTAMOV, Farukh. *Jupyter Notebook: "data$_e$xploring"* [online]. 2025. Available also from: `https://gitlab.fit.cvut.cz/rustafar/astronomical_data_ml/-/blob/main/SDSS/data_exploring.ipynb`. [accessed 2025-04-22].

24. 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`.

25. 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`.

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

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

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

29. 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].

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. 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.

32. LINTOTT, Chris J; SCHAWINSKI, Kevin; SLOSAR, Anže; LAND, Kate; BAMFORD, Steven; THOMAS, Daniel; RADDICK, M Jordan; NICHOL, Robert C; SZALAY, Alex; ANDREESCU, Dan, et al. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*. 2008, vol. 389, no. 3, pp. 1179–1189.

33. BAMFORD, Steven P; NICHOL, Robert C; BALDRY, Ivan K; LAND, Kate; LINTOTT, Chris J; SCHAWINSKI, Kevin; SLOSAR, Anže; SZALAY, Alexander S; THOMAS, Daniel; TORKI, Mehri, et al. Galaxy Zoo: the dependence of morphology and colour on environment. *Monthly Notices of the Royal Astronomical Society*. 2009, vol. 393, no. 4, pp. 1324–1352.

34. MACLEOD, CL; BROOKS, K; IVEZIĆ, Ž; KOCHANEK, CS; GIBSON, R; MEISNER, A; KOZŁOWSKI, S; SESAR, B; BECKER, AC; DE VRIES, WH. Quasar selection based on photometric variability. *The Astrophysical Journal*. 2011, vol. 728, no. 1, p. 26.

35. PETRILLO, Carlo Enrico; TORTORA, Crescenzo; CHATTERJEE, S; VERNARDOS, G; KOOPMANS, LVE; VERDOES KLEIJN, G; NAPOLITANO, Nicola Rosario; COVONE, G; SCHNEIDER, P; GRADO, Aniello, et al. Finding strong gravitational lenses in the kilo degree survey with convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*. 2017, vol. 472, no. 1, pp. 1129–1150.

36. JACOBS, Colin; COLLETT, Thomas; GLAZEBROOK, K; MCCARTHY, C; QIN, AK; ABBOTT, TMC; ABDALLA, FB; ANNIS, J; AVILA, S; BECHTOL, K, et al. Finding high-redshift strong lenses in DES using convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*. 2019, vol. 484, no. 4, pp. 5330–5349.

37. LOCHNER, Michelle; MCEWEN, Jason D; PEIRIS, Hiranya V; LAHAV, Ofer; WINTER, Max K. Photometric supernova classification with machine learning. *The Astrophysical Journal Supplement Series*. 2016, vol. 225, no. 2, p. 31.

38. HOSSEINZADEH, Griffin; DAUPHIN, Frederick; VILLAR, V Ashley; BERGER, Edo; JONES, David O; CHALLIS, Peter; CHORNOCK, Ryan; DROUT, Maria R; FOLEY, Ryan J; KIRSHNER, Robert P, et al. Photometric classification of 2315 Pan-STARRS1 supernovae with Superphot. *The Astrophysical Journal*. 2020, vol. 905, no. 2, p. 93.

39. SMOLINSKI, Jason P; MARTELL, Sarah L; BEERS, Timothy C; LEE, Young Sun. A Survey of CN and CH Variations in Galactic Globular Clusters from SDSS Spectroscopy. *arXiv preprint arXiv:1105.5378*. 2011.

40. NESS, Melissa; HOGG, David W; RIX, H-W; HO, Anna YQ; ZASOWSKI, Gail. The cannon: A data-driven approach to stellar label determination. *The Astrophysical Journal*. 2015, vol. 808, no. 1, p. 16.

41. PENG, Ying-jie; LILLY, Simon J; KOVAČ, Katarina; BOLZONELLA, Micol; POZZETTI, Lucia; RENZINI, Alvio; ZAMORANI, Gianni; ILBERT, Olivier; KNOBEL, Christian; IOVINO, Angela, et al. Mass and Environment as Drivers of Galaxy Evolution in SDSS and zCOSMOS and the Origin of the Schechter Function. *The Astrophysical Journal*. 2010, vol. 721, no. 1, p. 193.

42. WETZEL, Andrew R; TINKER, Jeremy L; CONROY, Charlie; VAN DEN BOSCH, Frank C. Galaxy evolution in groups and clusters: satellite star formation histories and quenching time-scales in a hierarchical Universe. *Monthly Notices of the Royal Astronomical Society*. 2013, vol. 432, no. 1, pp. 336–358.

43. PENG, Min. *| Fusion model. Left: Late Fusion. Middle: Early Fusion. Right: Slow... | Download Scientific Diagram* [online]. [N.d.]. Available also from: `https://www.researchgate.net/figure/Fusion-model-Left-Late-Fusion-Middle-Early-Fusion-Right-Slow-Fusion-The-red_fig6_320395034`. [accessed 2025-05-15].

44. 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].

45. DAVIS, Steven; MERMELSTEIN, Paul. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*. 1980, vol. 28, no. 4, pp. 357–366.

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

47. 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.

48. 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.

49. 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.

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

51. 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.

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

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

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

55. 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.

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

57. 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.

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

59. *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].

60. 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].

61. *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].

62. 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.

Directory Structure

- images .... directory of image assets
  - data .... common theoretical images
  - data_explore_photo .... images for the Data Exploration chapter
  - ml_photos .... images for ML-related chapters
  - qual_metrics .... ML results comparisons across resolutions
- text .... LaTeX source files
  - text.tex .... main chapter content
  - bib-database.bib .... bibliography database
  - medium.tex .... directory tree stored here
- src .... source code and notebooks
  - SDSS .... notebooks and analysis for SDSS dataset
    - Dimensionality_reduction .... experiments on dimensionality reduction
    - ML_Q[0] .... ML on $Q = 0$ resolution
    - MI_Q[2] .... Mutual information analysis at $Q = 2$
    - ML_qualities .... ML experiments evaluating quality metrics
  - Scene .... scene-related code and assets
- rustafar-assignment.pdf .... thesis assignment PDF
- ctufit-thesis.tex .... main thesis driver
- ctufit-thesis.cls .... CTU FIT thesis class file
- creationdate.timestamp .... template creation timestamp
- changelog.md .... template changelog
- LICENSE .... license file
- thesis.pdf .... compiled thesis PDF