Bachelor's thesis

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

Bc. Farukh Rustamov

Faculty of Information Technology
Department of Applied Mathematics
Supervisor: ————
April 30, 2025

FACULTY OF INFORMATION TECHNOLOGY CTU IN PRAGUE

Assignment of bachelor's thesis

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

Instructions

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

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

The main tasks are:

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

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

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

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

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

I would like to express my sincere gratitude to my supervisor, ***RNDr. Petr Škoda, CSc.****, for his valuable guidance, insightful feedback, and continuous support throughout the development of this thesis.*
I would also like to thank ***Ing. Ondřej Podsztavek*** *for his expert advice and assistance with machine learning methods, which significantly contributed to the quality and depth of the experimental work.*
All computations for this thesis were carried out on the RCI cluster, providing access to high-performance computing resources and enabling more complex and large-scale machine learning experiments. The authors acknowledge the support of the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics".
The access to the computational infrastructure of the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics" is also gratefully acknowledged. Most of the experiments and data processing were carried out using the RCI cluster.
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. SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration, including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatório Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt Uni-

Declaration

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

In Prague on April 30, 2025

Abstract

In this thesis, we present a comprehensive study on predicting the star formation rate (SFR) in galaxies using multimodal data from the Sloan Digital Sky Survey (SDSS, DR7). We begin by filtering the original SFR catalogue to a high-quality subset of 11 179 galaxies with valid five-band photometry and one-dimensional spectra, processed via the HiSS-Cube pipeline to generate multi-resolution image cutouts (64×64 to 4×4 px) and spectral samplings (4620 to 289 bins) while preserving measurement uncertainties.

We evaluate three classes of regression models—Decision Tree (DT), VGGNet12 convolutional neural network, and LightGBM gradient boosting—under three modalities: photometry-only, spectroscopy-only, and multimodal fusion (early and late fusion). Hyperparameter tuning is performed via grid search and five-fold cross-validation, and model performance is assessed by R^2, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Normalized Median Absolute Deviation (NMAD).

The best results are achieved by the early-fusion LightGBM model, reaching $R^2 = 0.308$, MAE=0.19, RMSE=0.32, demonstrating the strength of tree-based learners in combining visual and spectral features. VGGNet12 on photometric images alone also performs strongly ($R^2 = 0.262$), highlighting the power of deep CNNs for morphological analysis. Notably, the lowest spectral resolution often yielded better generalization due to implicit noise smoothing.

Our findings confirm that multimodal machine learning can effectively capture complementary astrophysical cues for accurate SFR estimation. The methodological framework laid out here paves the way for exploring advanced fusion techniques (e.g., attention-based models) and incorporating additional data modalities such as environmental or kinematic measurements in future research.

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

Abstrakt

V této diplomové práci jsme se zaměřili na predikci rychlosti formování hvězd (SFR) v galaxiích využitím multimodálních dat ze Sloan Digital Sky Survey (SDSS). Nejprve jsme z původního katalogu SFR (DR7) vyfiltrovali objekty s chybějícími nebo nekvalitními hodnotami, čímž vznikla konečná sada 11 179 galaxií s validními fotometrickými i spektroskopickými měřeními. Pro předzpracování dat jsme využili HiSS-Cube pipeline, která generuje více rozlišení obrazových výřezů (64×64 až 4×4 px) a spekter (4620 až 289 vzorků), přičemž zachovává nejistoty měření a umožňuje efektivní dotazování.

Pro regresní predikci logaritmu SFR (`AVG`) jsme porovnávali tři třídy modelů: rozhodovací stromy (DT), konvoluční neuronové sítě (VGGNet12) a gradientní boostování (LightGBM). Každý model jsme testovali ve třech režimech: fotografie-only, spektra-only a multimodální fúze (early fusion i late fusion). Hyperparametry jsme ladili pomocí grid-search a 5-násobné křížové validace. Jako metriky jsme sledovali R^2, MAE, RMSE a NMAD.

Výsledky ukázaly, že nejlepší výkon dosáhl early-fusion model LightGBM ($R^2 = 0.308$, MAE=0.19, RMSE=0.32) díky schopnosti stromových modelů efektivně kombinovat vizuální a spektrální rysy. VGGNet12 na obrázcích také dosahuje vysoké kvality ($R^2 = 0.262$), což potvrzuje sílu hlubokých CNN pro extrakci morfologických ukazatelů. Zajímavě nejnižší rozlišení spekter poskytlo lepší generalizaci díky účinku vestavěného vyhlazování.

Tato práce demonstruje, že multimodální přístupy dokáží zachytit komplexní fyzikální a morfologické informace klíčové pro odhad SFR a otevírá cestu k dalšímu zkoumání pokročilých fúzních technik či zapojení doplňkových dat (kinematika, prostředí).

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

Contents

List of Figures

List of Tables

List of code listings

List of abbreviations

SDSS	Sloan Digital Sky Survey
SFR	Star Formation Rate
CNN	Convolutional Neural Network
MFCC	Mel-Frequency Cepstral Coefficients
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
NMAD	Normalized Median Absolute Deviation
DT	Decision Tree
VGG	Visual Geometry Group
ML	Machine Learning
LLM	Large Language Model
LightGBM	Light Gradient Boosting Machine
HDF5	Hierarchical Data Format version 5
RCI	Research Computing Infrastructure
MLP	Multilayer Perceptron
PCA	Principal Component Analysis
t-SNE	t-Distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
VO	Virtual Observatory
HMS	Hierarchical Semi-Sparse

Chapter 1

Introduction

1.1 General Description and Relevance of the Study

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

In astrophysics, large-scale surveys like the Sloan Digital Sky Survey (SDSS) [1] provide both photometric and spectroscopic data for millions of celestial objects. These complementary modalities offer unique views: images capture structural and morphological features, while spectra encode detailed physical and chemical properties.

This thesis investigates the application of multimodal machine learning techniques to predict the **star formation rate (SFR)** [2] in galaxies using data from SDSS. The motivation lies in the need to efficiently process massive astronomical datasets and build models that leverage the strengths of both image-based and spectroscopic inputs.

1.2 SDSS Data 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 8000 deg^2 with more than 1.6 million galaxy spectra [3]. 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 [4].

In this thesis we primarily use data from SDSS Data Release 7 (DR7) [5]. 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.

1.3 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, we developed the *Hierarchical Semi-Sparse Cube (HiSS-Cube)* framework based on HDF5, which offers:

- **Multi-domain fusion:** Supports imaging, spectral, and time-series data in a single 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 changing model input sizes.
- **Virtual Observatory compatibility:** Exports to VOTable/FITS for use in standard VO tools.
- **Performance gains:** Benchmarks on SDSS Stripe 82 data show HiSS-Cube is faster by orders of magnitude compared to raw FITS exports for both visualization and ML pipelines.

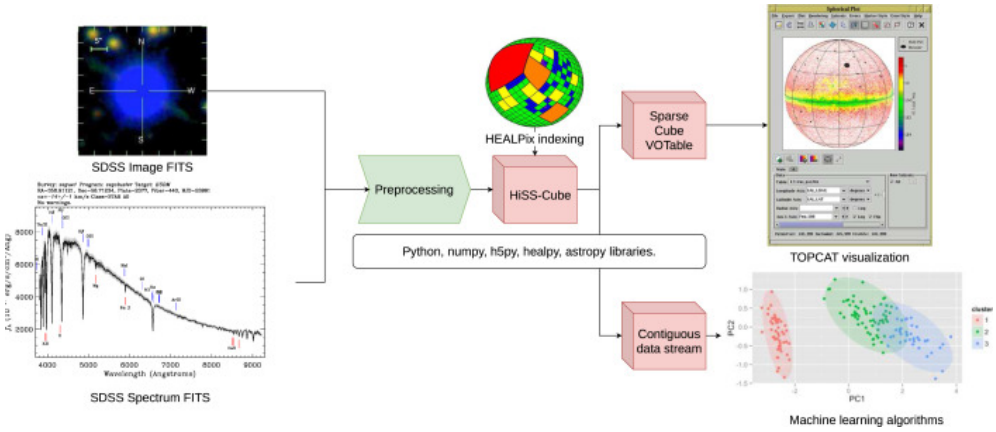


Figure 1.1 HiSS-Cube data flow pipeline: from SDSS raw FITS files to multi-layered, semi-transparent cubes in HDF5 for visualization and machine learning. Image from [6].

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. [6]

1.3.1 Prediction Experiments

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

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

1.3.2 Role of Spectroscopy vs. Photometry

Spectroscopic data provide direct physical diagnostics—emission-line luminosities (e.g., Hα) which scale with instantaneous SFR, as well as redshift measurements for distance correction [7]. Photometric images encode morphological details, color gradients, and integrated broadband flux, reflecting the galaxy's stellar population and dust content. By fusing these complementary views, our models can leverage both fine-scale spectral physics and global structural cues, leading to more robust and accurate SFR predictions.

1.4 Research Challenges

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

1. **Data Filtering.** The SDSS SFR catalog originally contains over 4.8 million entries, but only a fraction have both reliable multi-band cutouts and valid SFR measurements. We must exclude objects with missing photometry or spectroscopy, undefined SFR values (NaN or the placeholder -99), and non-galactic sources, reducing the sample to a few $\times 10^4$ galaxies suitable for regression [8].

2. **Quality of Images and Spectra.** The HiSS-Cube pipeline provides four image resolutions (64×64, 32x32, 16x16, 8×8 px) and four spectral samplings (4620,2310,1155,577,289 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 [6].

3. **Multiple Objects in One Image.** SDSS cutouts sometimes include overlapping galaxies or stars, leading to blended light profiles that confuse downstream feature extractors. To ensure each input represents a single target galaxy, we apply automatic segmentation via thresholding and connected-component labeling, flagging and removing multi-object cutouts [9, 10].

1.5 Objectives and Tasks

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

1. Perform a detailed analysis of the raw data, assess its quality, and apply filtering.

2. Develop algorithms for the automatic detection and isolation of objects within images.

3. Investigate the impact of different quality levels of images and spectra on prediction accuracy.

4. Compare the effectiveness of models using single modalities with multimodal approaches.

5. Conduct a comparative study on the publicly available Scene dataset, adapting insights to SDSS in order to validate our multimodal pipeline under controlled conditions.

6. Quantify the relative performance gain of multimodal fusion over unimodal (image-only and spectrum-only) baselines on a structurally similar external dataset to demonstrate the added value of combining modalities.

7. Benchmark and compare training and inference runtimes of all models and modalities on both SDSS and the external dataset, to assess computational scalability and guide practical deployment strategies.

1.6 Terminology and Illustrations

1.6.1 Spectra and Spectral Analysis

1.6.1.1 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 [11].

1.6.1.2 Why Spectral Analysis Is Needed

- **Chemical Composition:** Spectral lines from elements such as hydrogen, oxygen, nitrogen, and iron appear at characteristic wavelengths, and their relative intensities allow us to derive abundances and metallicity in the interstellar medium. For example, the ratio of [O III] to H β lines is a common metallicity diagnostic [12]. These abundance measurements are crucial for understanding galactic chemical evolution and enrichment histories [11].

- **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 [12]. 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 [11]. Understanding these conditions informs models of star-formation efficiency and feedback processes.

All of these diagnostics are discussed in [11, p. 1–6].

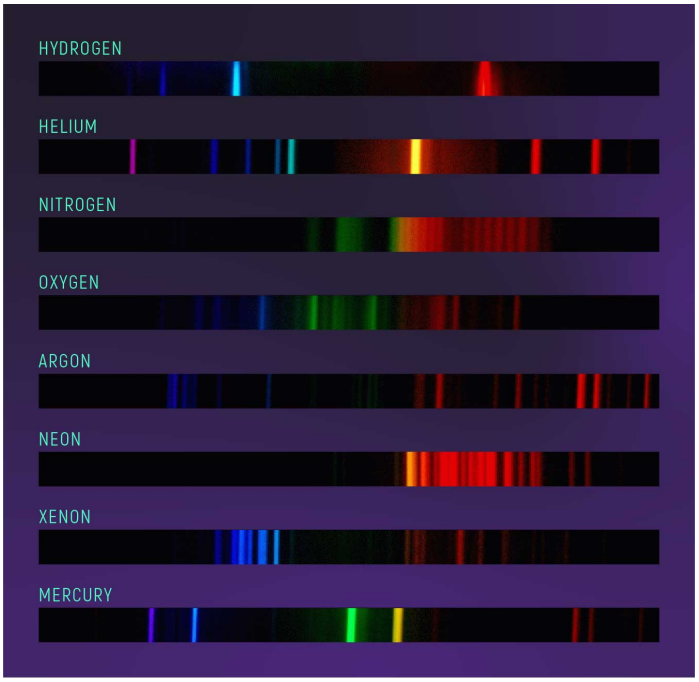


Figure 1.2 Example of atomic spectral lines for different elements.[13]

1.6.2 The SDSS *u, g, r, i, z* Filters

SDSS uses five broadband filters—u, g, r, i, and z—with effective wavelengths of $u = 354\,\text{nm}$, $g = 477\,\text{nm}$, $r = 623\,\text{nm}$, $i = 762\,\text{nm}$, and $z = 913\,\text{nm}$. Their full-width at half-maximum (FWHM) bandwidths are approximately $\Delta u \approx 56\,\text{nm}$, $\Delta g \approx 138\,\text{nm}$, $\Delta r \approx 138\,\text{nm}$, $\Delta i \approx 152\,\text{nm}$, and $\Delta z \approx 95\,\text{nm}$ [14].

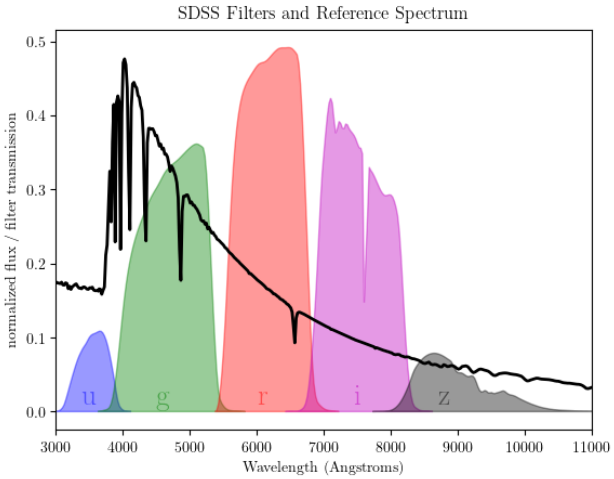


Figure 1.3 Transmission curves of the SDSS *u, g, r, i, z* filters.

1.6.3 Star Formation Rate (SFR)

1.6.3.1 What Is 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 [15]. 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 [16].

1.6.3.2 How SFR Is Determined

Because stars of different masses and ages radiate energy differently, astronomers use several complementary tracers to estimate SFR:

- **Hydrogen Emission Lines.** Massive young stars emit ultraviolet light that ionizes hydrogen. When the gas recombines, it produces lines like H α and H β. After correcting for dust and aperture losses, the H α luminosity relates to SFR as [7]:

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

- **Ultraviolet Continuum.** The UV light (e.g. at 1500 Å) traces stars formed over the last ∼100 Myr. It is calibrated via [17]:

$$\mathrm{SFR}\,(M_\odot\,\mathrm{yr}^{-1}) \approx 1.4 \times 10^{-28}\, L_\nu\ (\mathrm{erg\,s}^{-1}\,\mathrm{Hz}^{-1}),$$

  though dust extinction can introduce uncertainties.

- **Infrared Emission.** Dust absorbs UV/optical light and re-emits it in the far-infrared. The total IR luminosity compensates for obscured star formation [17]:

$$\mathrm{SFR}\,(M_\odot\,\mathrm{yr}^{-1}) \approx 4.5 \times 10^{-44}\, L_{\mathrm{TIR}}\ (\mathrm{erg\,s}^{-1}).$$

- **Hybrid Indicators.** To capture both unobscured and dust-hidden stars, one often combines H α (or UV) with mid-infrared (e.g. 24 µm):

$$\mathrm{SFR} \approx 7.9 \times 10^{-42}\, L(\mathrm{H}\alpha)_{\mathrm{obs}} + 0.031\, L(24\,\mu\mathrm{m}),$$

  which reduces systematic bias to $\sim 30\%$ [18].

- **Radio Continuum.** At $\sim$ 1.4 GHz, non-thermal synchrotron emission from supernova remnants provides an extinction-free SFR estimate over $\sim$100 Myr:

$$\mathrm{SFR}\,(M_\odot\,\mathrm{yr}^{-1}) \approx 1.0 \times 10^{-28}\, L_\nu(1.4\,\mathrm{GHz})\ (\mathrm{erg\,s^{-1}\,Hz^{-1}}),$$

  with uncertainties $\lesssim 20\%$ [19].

1.6.3.3 Why SFR Is a Key Galactic Parameter

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

- **Stellar Mass Assembly.** The SFR directly measures the conversion rate of cold gas into stars, driving the build-up of stellar mass and shaping the galaxy stellar mass function over cosmic time [17].

- **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 [16].

Chapter 2

Data Exploration

2.1 Dataset Overview and Initial Filtering

We source our sample from the SDSS Data Release 7 star formation rate (SFR) catalog, which initially contains 4 851 200 objects. To ensure that every galaxy has both imaging and spectroscopic data, we retain only those entries with available multi-band cutouts and 1D spectra, reducing the sample to 151 190 records. Next, we remove entries where the logarithmic SFR indicator `AVG` is undefined (NaN), leaving 34 613 objects. Finally, we exclude the placeholder value `AVG` $= -99$, resulting in 30 752 records. Of these, 16 841 have `FLAG`=0 (high-quality SFR estimates) and 13 911 have `FLAG`$\neq$0 [8, 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`$\neq$0)	13 911

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

Because we leverage the HiSS-Cube framework—a scalable pipeline for hierarchical semi-sparse cubes that preserves measurement uncertainties and precomputes cutouts—each galaxy in our high-quality subset is accompanied by five image quality levels and five spectral resolutions [6]. Moreover, each of these variants carries the same `AVG` SFR label, simplifying our supervised

learning setup.

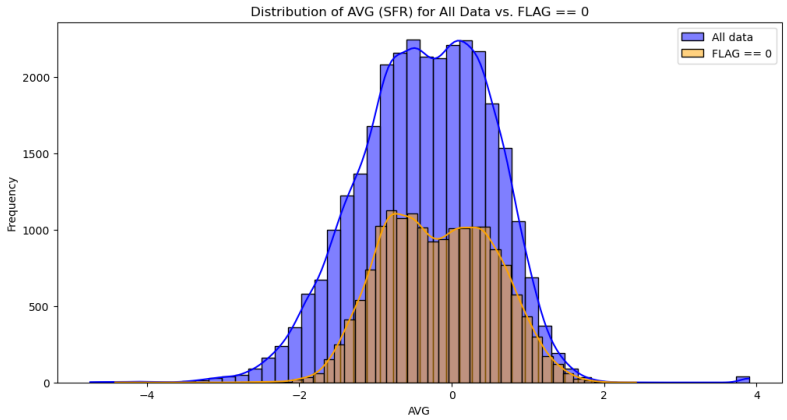


Figure 2.1 Distribution of `AVG` ($\log_{10}$ SFR) in the filtered sample.

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 0.

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) [14]. Each image captures a specific portion of the spectrum, enabling a detailed analysis of the structural and physical properties of the objects.
- **Spectroscopic Data.** In addition to the photometric images, each object is provided with a spectrum that offers information on its chemical composition, temperature, and dynamics.

2.3 Image and Spectrum Data Availability

Thanks to the HiSS-Cube pipeline [6], each high-quality galaxy (`FLAG`=0) is preprocessed into a multi-resolution "cube" that preserves uncertainties. For our regression experiments, we retrieve five image resolutions and five spectral samplings per object (see Fig. **??**).

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

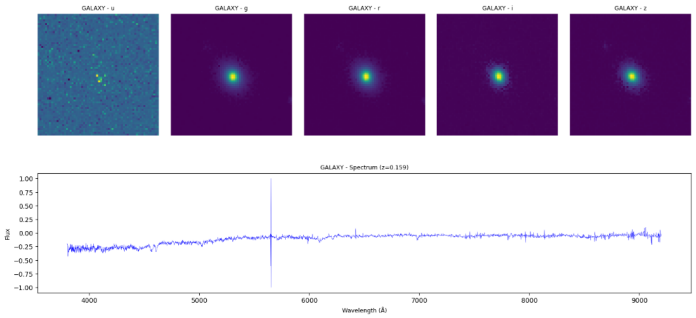


Figure 2.2 An example of an object. Top 5 pixel photos, bottom a spectrum.

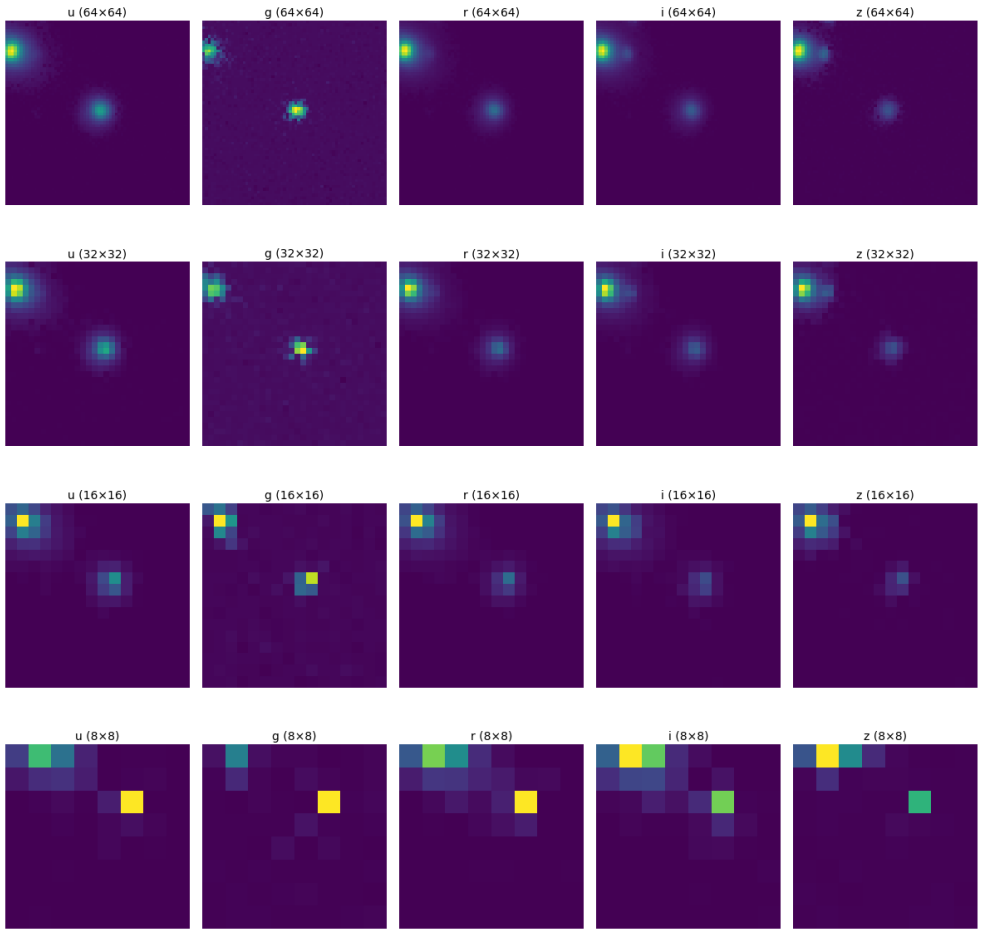


Figure 2.3 HiSS-Cube image outputs for a single galaxy at five resolution levels (64×64 to 4×4 pixels).

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

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

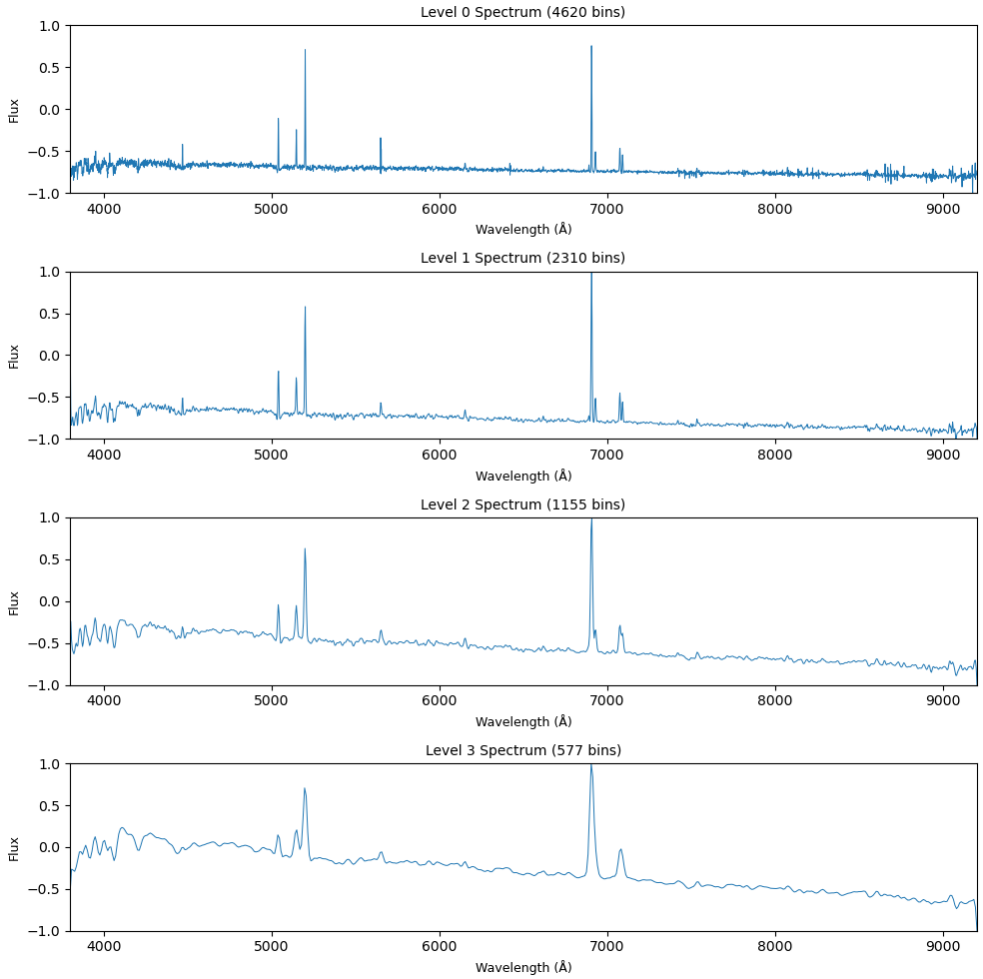


Figure 2.4 HiSS-Cube spectral outputs for the same galaxy at five sampling levels (4620 to 289 bins).

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

2.4 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." [8]

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

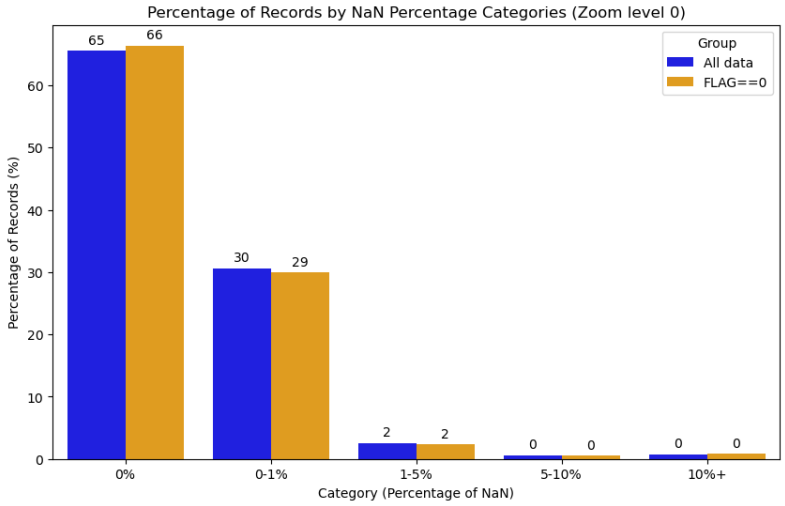


Figure 2.5 Percentage of records by NaN percentage categories at Zoom level 0, comparing all data vs. `FLAG`=0 subset.

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

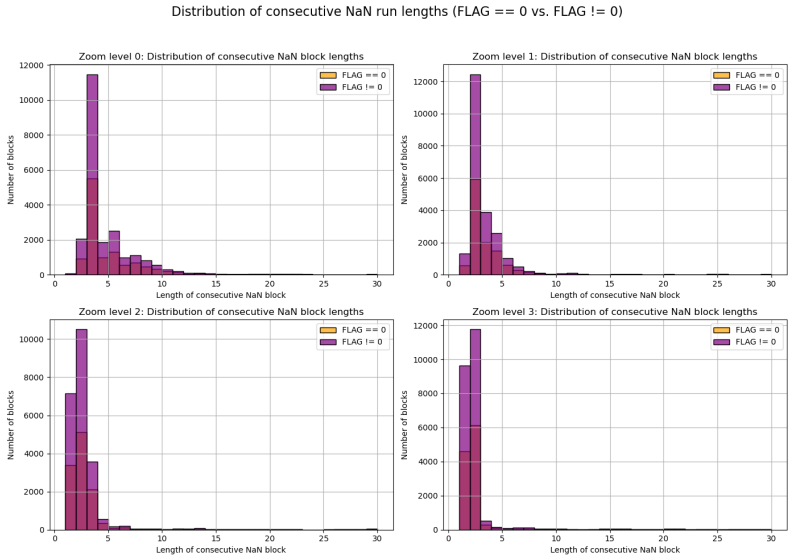


Figure 2.6 Distribution of consecutive NaN run lengths at each resolution for `FLAG`=0.

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.

2.5.4 NaN Occurrence Along Wavelength

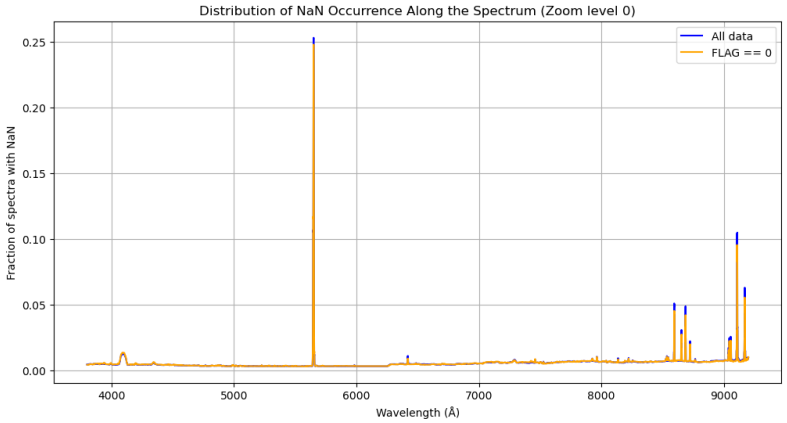
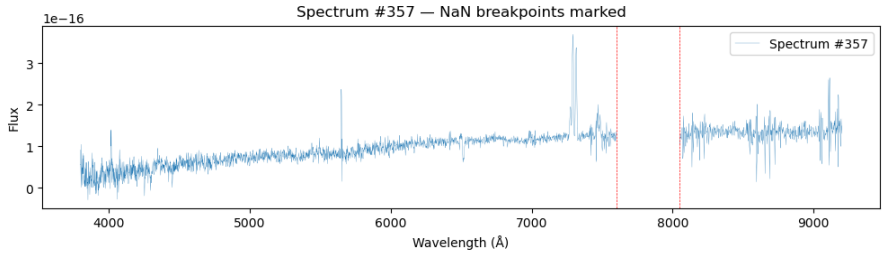


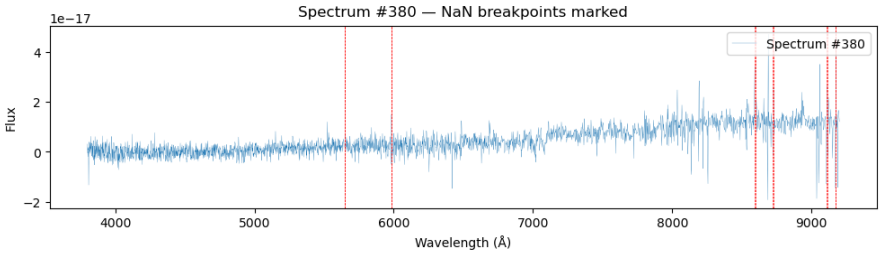
Figure 2.7 Typical wavelength regions where NaN gaps commonly occur (Zoom level 0).

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; notably, there is no wavelength where 0% of spectra are missing data, indicating that every channel is affected by occasional dropouts or quality flags. The sharp spike at $\sim 5500\,\text{Å}$ coincides with the dichroic split between the blue and red arms of the SDSS spectrograph, where stitching mismatches and calibration uncertainties often lead to flagged pixels [24]. The elevated NaN occurrence near $\sim 9000\,\text{Å}$ arises from the declining quantum efficiency of the red CCDs and strong telluric emission lines (e.g. atmospheric OH), which reduce the signal-to-noise ratio and trigger data quality filters [25].



(a) Spectrum 357 from SDSS showing regions of missing values (NaN). Red vertical dashed lines indicate the wavelength ranges (around 7400 Å and 8000 Å) where data are absent.



(b) Spectrum 380 from SDSS with regions of missing data (NaN) highlighted. Red vertical dashed lines mark the wavelengths—around 5600 Å, 6000 Å, 8850 Å, 9050 Å, and 9200 Å—where flux values are absent.

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

2.6 Detection and Removal of Multi-Object Cutouts

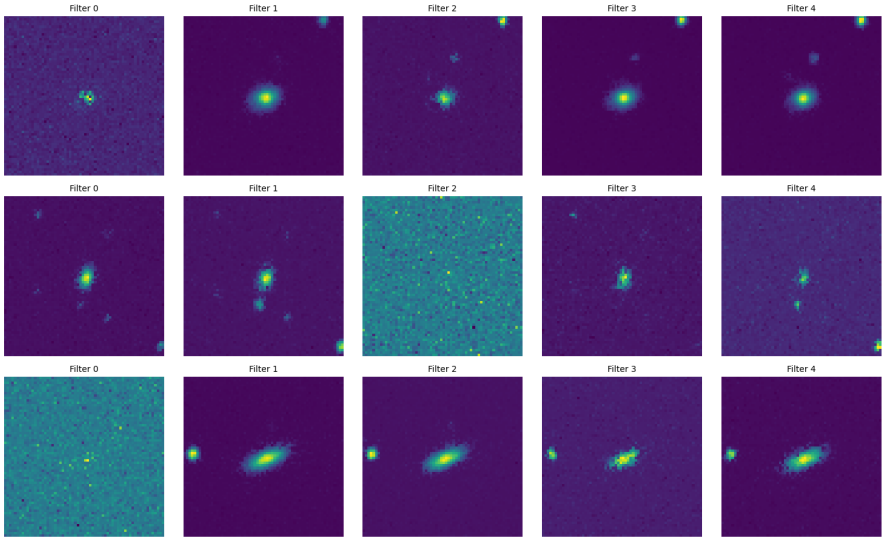


Figure 2.9 Example of a cutout containing multiple detected sources, excluded from the final sample [23] .

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 [9]. 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 [9] as well as the standard workflow described in Gonzalez and Woods’s digital image processing text [10].

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.
- **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.
- **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.

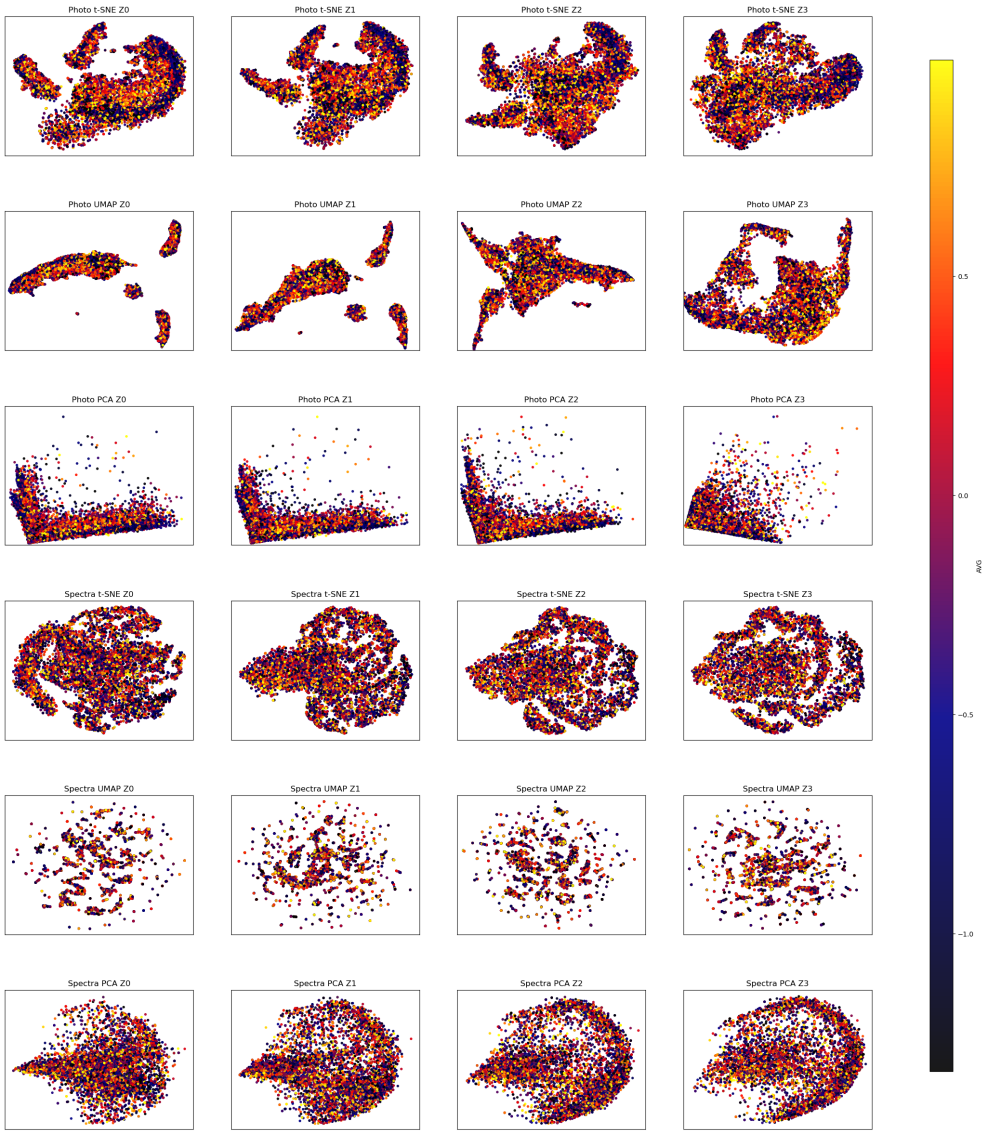


Figure 2.10 Embeddings of image and spectral data at four zoom levels (Z0–Z3) using t-SNE, UMAP, and PCA, colored by `AVG` [29].

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].

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

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 seeks to integrate and jointly reason over heterogeneous data sources—such as images, text, audio, and structured signals—to build models that capture complementary information and achieve higher performance than any single modality alone [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 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. [32]

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. [32]

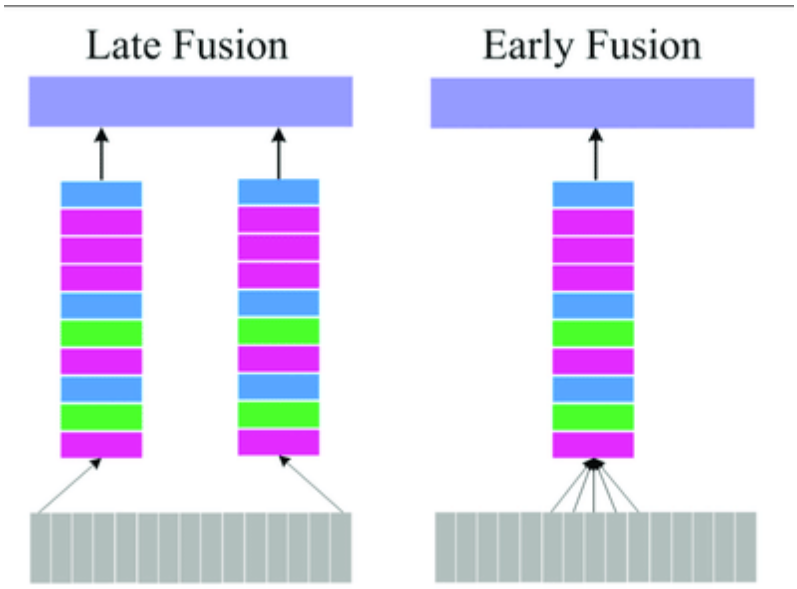


Figure 3.1 Illustration of Late and Early Fusion strategies in multimodal learning [32].

3.2.1 Suitability of SFR Prediction as a Regression Task

Predicting the star formation rate (SFR) is inherently a continuous regression problem rather than a classification task. The target variable, typically expressed as $\log_{10}(\mathrm{SFR}, /, M_{\odot}, \mathrm{yr}^{-1})$, varies smoothly across galaxies of different morphologies and physical conditions. Multimodal fusion allows us to exploit both morphological cues from images and physical diagnostics from spectra to minimize prediction error in this continuous space.

3.3 Scene Dataset Example

To illustrate the general benefits of multimodal learning, we conducted preliminary experiments on the publicly available Scene dataset [33], which provides

two modalities for environmental scene classification:

- **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.

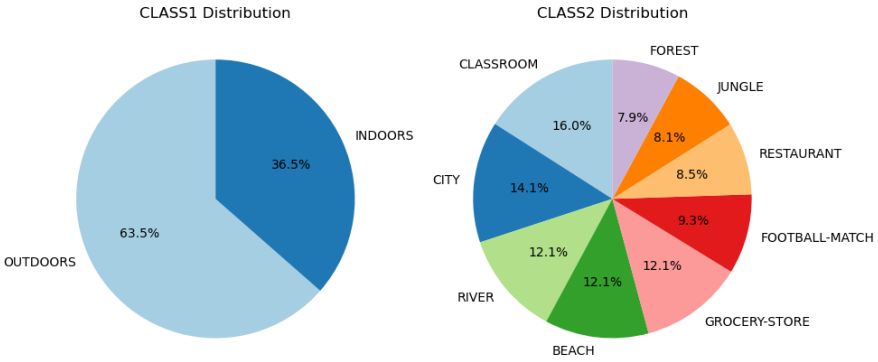


Figure 3.2 CLASS1 (left) and CLASS2 (right) label distributions for the Scene dataset.

Although both modalities individually yield high classification accuracy (¿99%), multimodal fusion (early or late) 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 Star–Galaxy–Quasar classification

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

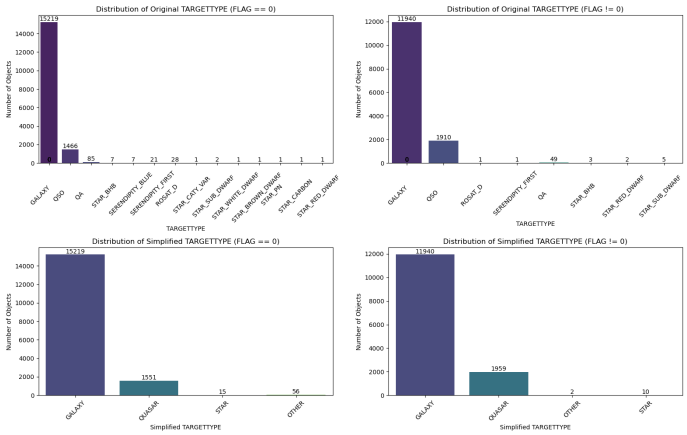


Figure 4.1 Class distribution for star–galaxy–quasar labels: galaxies outnumber quasars by a factor of 10, and stars comprise fewer than 30 objects [23].

4.2 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 [34].
- **Gradient Boosting Machine (LightGBM).** An efficient implementation of gradient-boosted decision trees optimized for speed and memory [35].

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 [36, 37].

4.3.2 Preprocessing

- *Images:* pixel values are linearly scaled to $[0, 1]$ by dividing by 255 [38], then flattened for decision-tree/LightGBM models or fed as 2D arrays into VGGNet12 [39].
- *Spectra:* Any object with NaN flux values removed, yielding 11,179 gap-free spectra[40].
- *Early Fusion:* Concatenate image and spectral vectors into one feature vector [41].
- *Late Fusion:* Average photo-only and spec-only model predictions[41].

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 [**Prechelt2002early**][42].

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 [43]. 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. Proper hyperparameter selection is crucial because under-regularized models overfit and under-fitted models underutilize the available signal.

4.3.4 Hyperparameter Tuning

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

VGGNet12: sweep over learning rate (`lr`) and fixed dropout=0.5, early stopping patience=30 [42, 44].

LightGBM: grid over `learning_rate` and `max_depth`, early stopping round=10 [45].

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].

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

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

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

- *Normalized Median Absolute Deviation (NMAD).* $1.4826 \times \mathrm{median}(|\epsilon - \mathrm{median}(\epsilon)|)$ [46].

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

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.

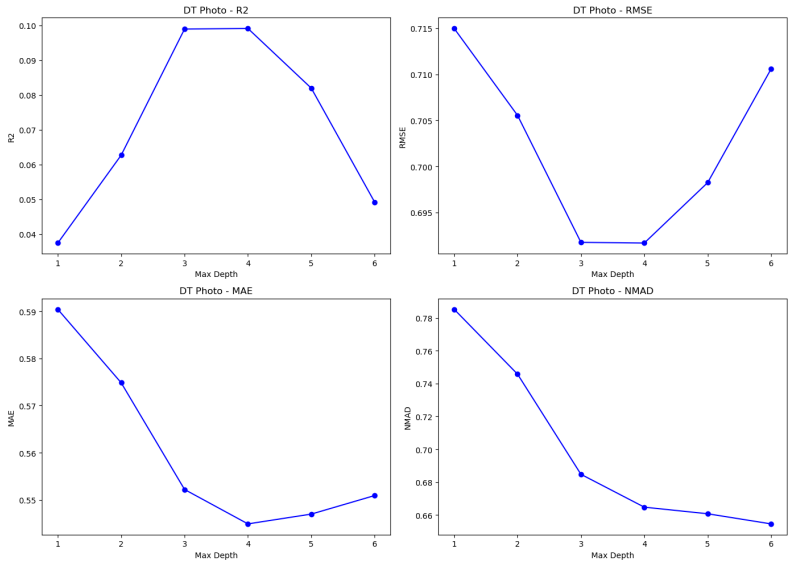


Figure 4.2 DT on photographs: R^2, MAE, RMSE, and NMAD vs. max. tree depth. Best $d = 4$ (all except NMAD).

Figure 1: Photo-only DT performance.

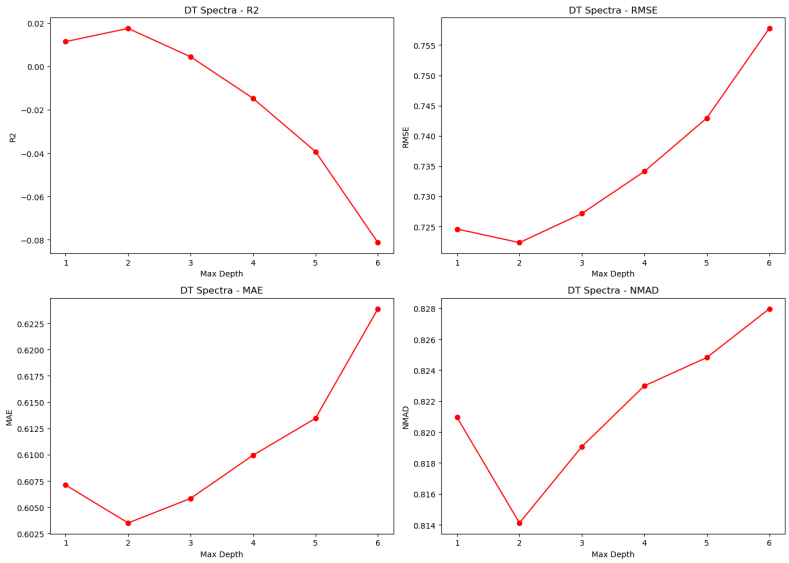


Figure 4.3 DT on spectra: R^2, MAE, RMSE, and NMAD vs. max. tree depth. Best $d = 2$.

Figure 2: Spectra-only DT performance.

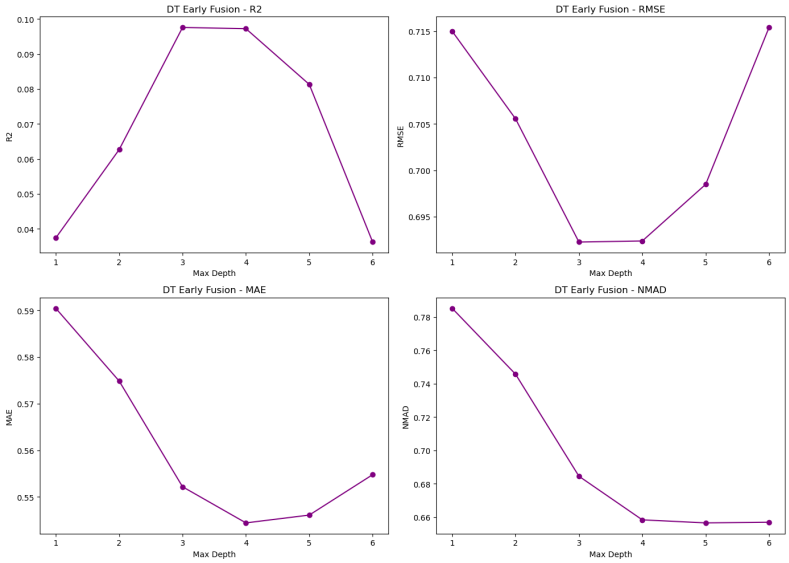


Figure 4.4 DT early fusion: R^2, MAE, RMSE, and NMAD vs. tree depth. Best $d = 3$ by R^2.

Figure 3: Early fusion DT performance.

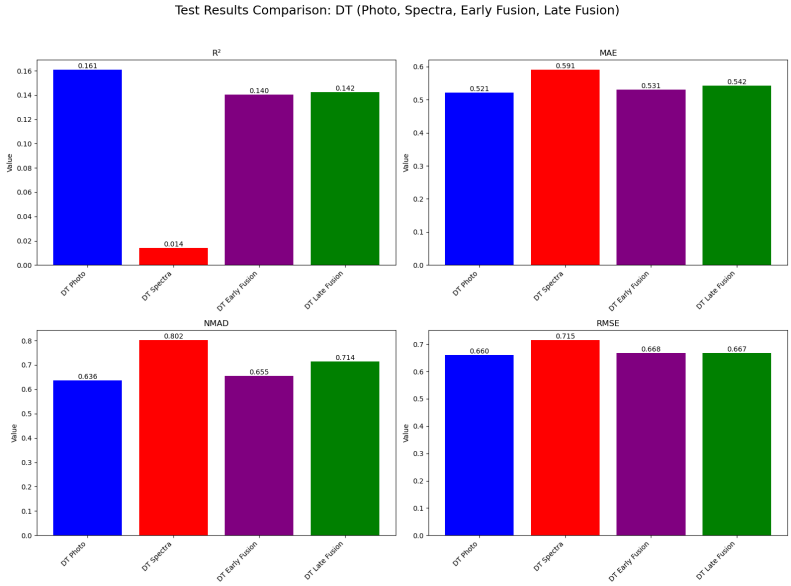


Figure 4.5 DT: metric comparison across modalities (photo, spectra, early, late).

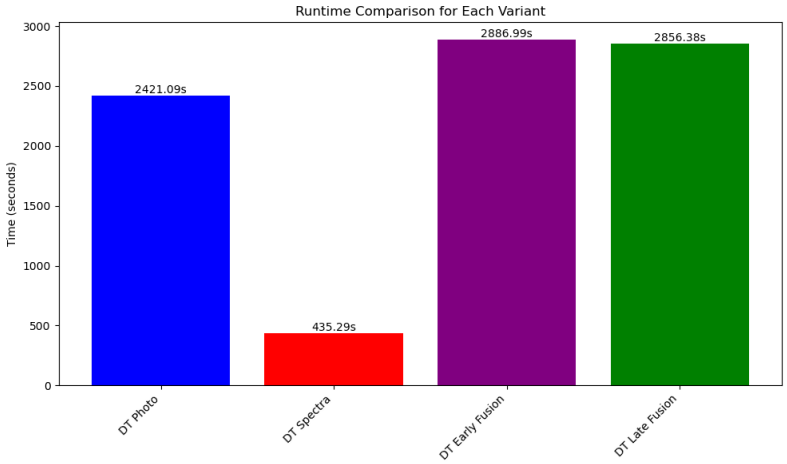


Figure 4.6 DT: wall-clock runtime across modalities.

4.6 Convolutional Neural Network: VGGNet12

The VGGNet12 model stacks 3×3 convolutions, max-pooling, then three FC layers with dropout, fine-tuned from ImageNet [34].

4.6.1 Architecture and Training Protocol

We optimize custom MSE loss,

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

using Adam, early stopping (patience=30), and focus hyperparameter tuning on learning rate [47, 44, 42].

4.6.2 Training Curves: Photographs

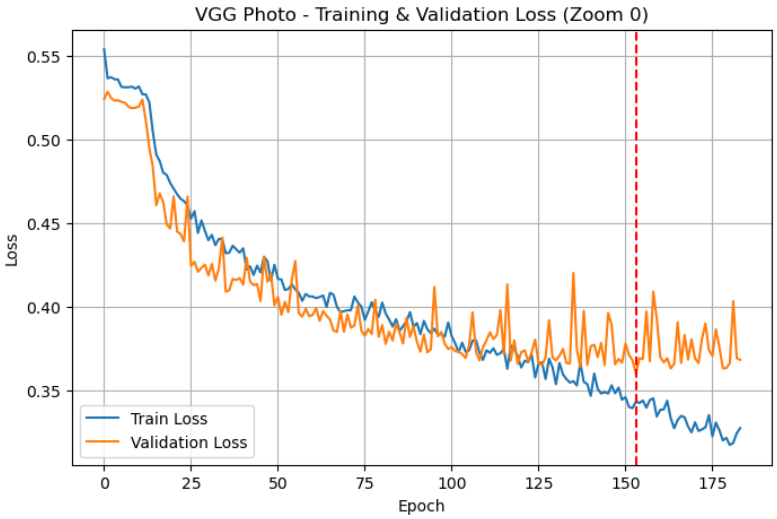


Figure 4.7 VGGNet12 photo: training (blue) vs. validation (orange) loss per epoch; red dashed line marks lowest val. loss.

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

4.6.3 Hyperparameter Sweep: Photographs

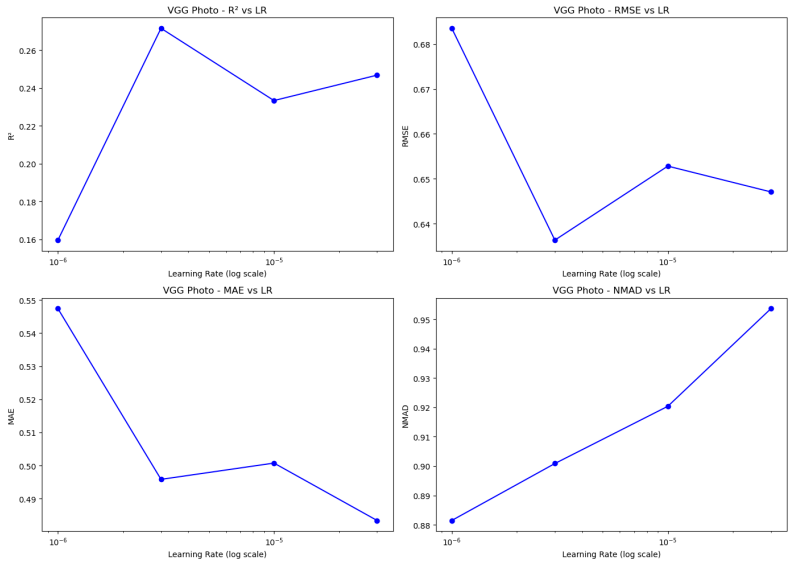


Figure 4.8 VGGNet12 photo: R^2, MAE, RMSE, NMAD vs. learning rate.

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

4.6.4 Training Curves: Spectra

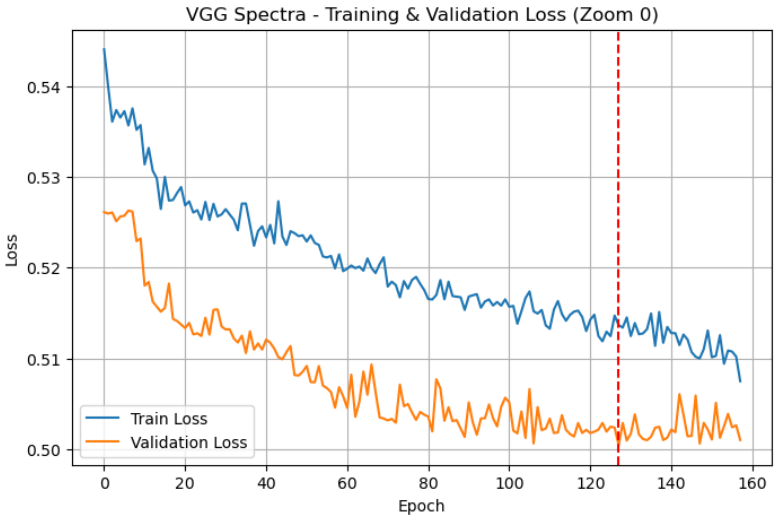


Figure 4.9 VGGNet12 spectra: training vs. validation loss per epoch; red dashed line = best epoch.

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

On this graph, it is particularly noticeable that there are epochs where the validation loss dips below the training loss. This behavior is expected in networks using dropout: during training dropout with $p = 0.5$ randomly deactivates neurons, adding noise and raising training loss, whereas no dropout is applied during validation, so the validation loss can occasionally be lower than the training loss [43].

4.6.5 Hyperparameter Sweep: Spectra

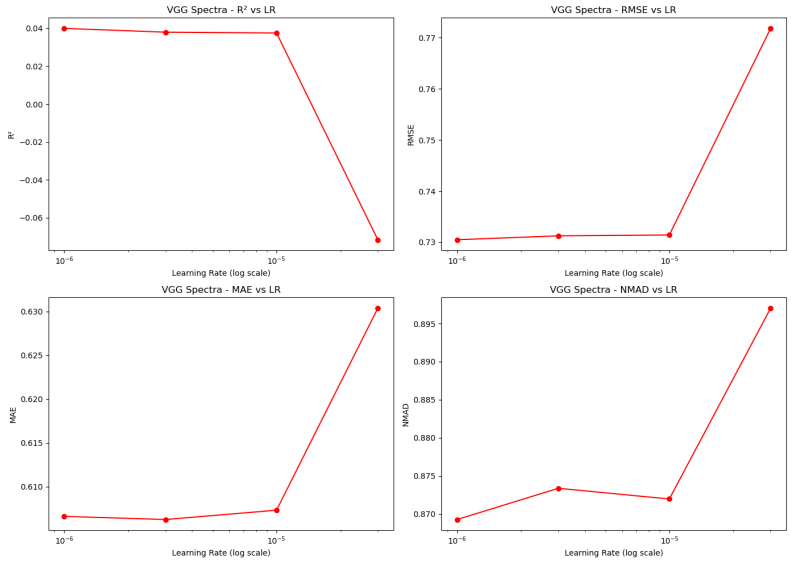


Figure 4.10 VGGNet12 spectra: R^2, MAE, RMSE, NMAD vs. learning rate.

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

4.6.6 Training Curves: Early Fusion

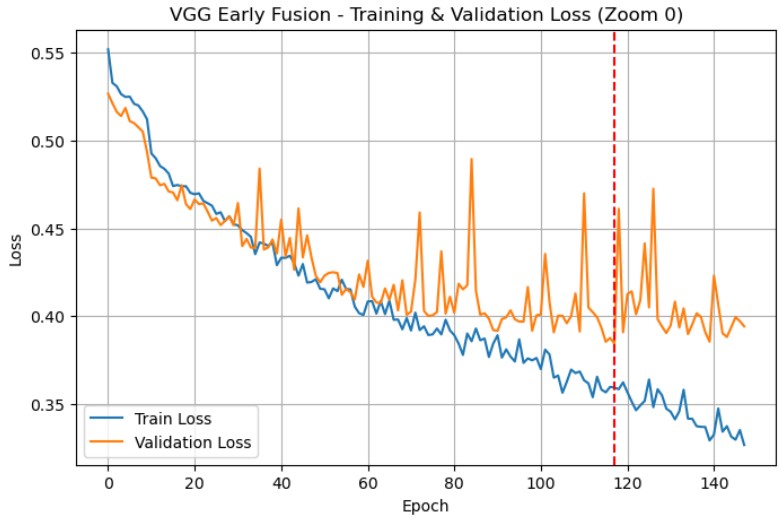


Figure 4.11 VGGNet12 early fusion: training vs. validation loss; red dashed line = best epoch.

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

4.6.7 Hyperparameter Sweep: Early Fusion

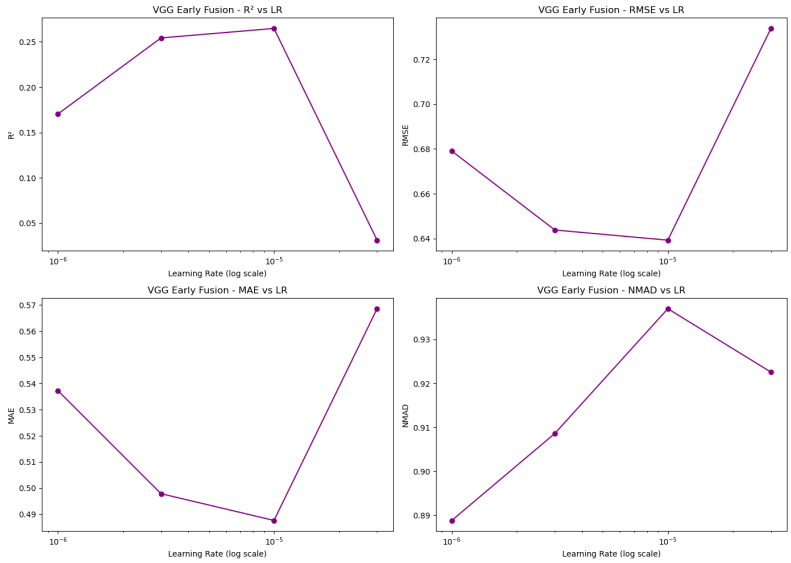


Figure 4.12 VGGNet12 early fusion: R^2, MAE, RMSE, NMAD vs. learning rate.

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

4.6.8 Overall Metrics and Runtime

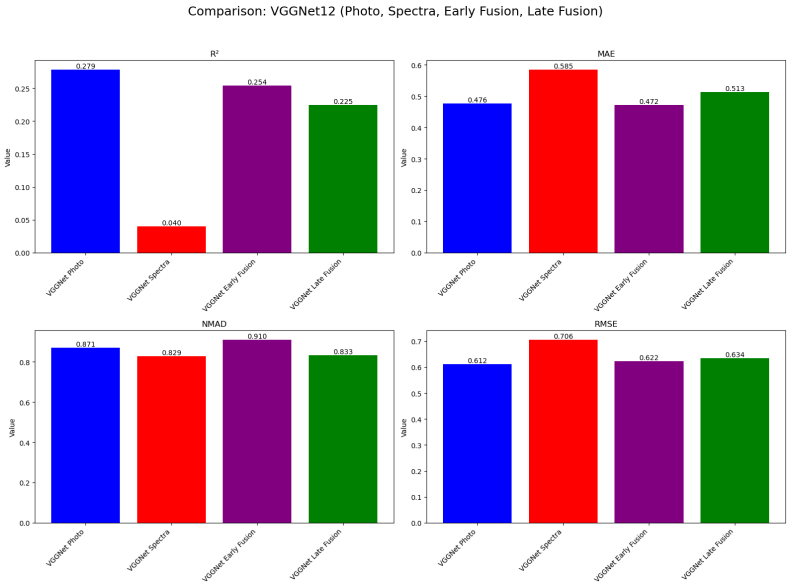


Figure 4.13 VGGNet12: metric comparison across modalities.

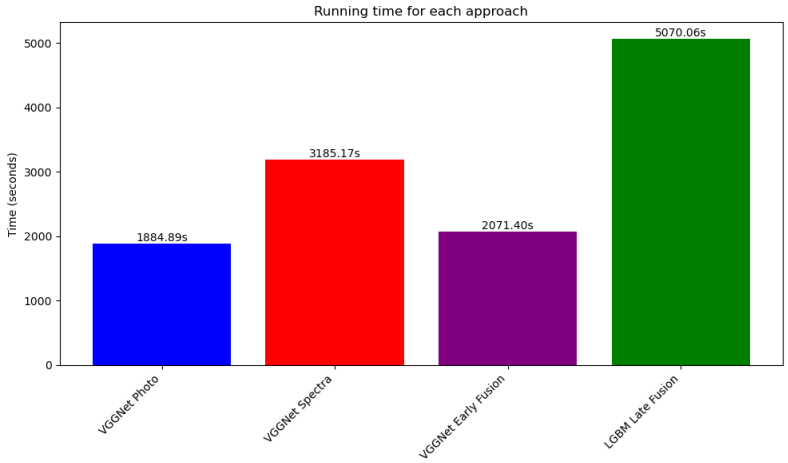


Figure 4.14 VGGNet12: wall-clock runtime across modalities.

4.7 Gradient Boosting Machine: LightGBM

LightGBM grows trees leaf-wise with histogram-based splitting and optimizes RMSE with early stopping (10 rounds) [35, 45].

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 [48].

4.7.2 Training Curves: Photographs

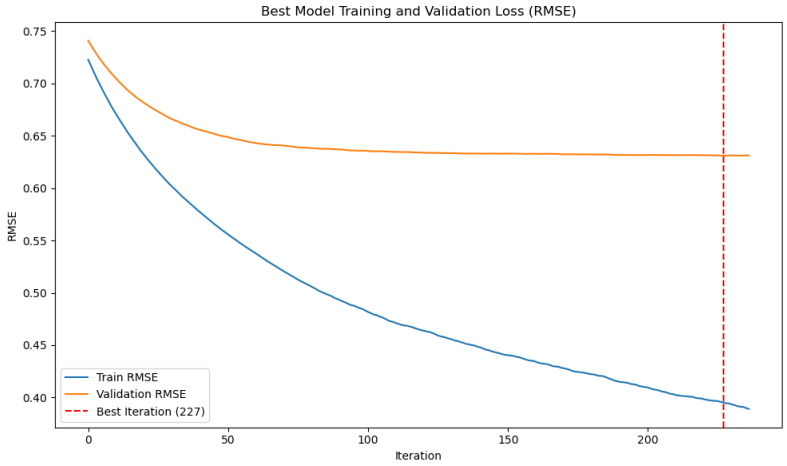


Figure 4.15 LightGBM photo: training vs. validation RMSE per iteration; red dashed line = best iteration.

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

4.7.3 Hyperparameter Sweep: Photographs

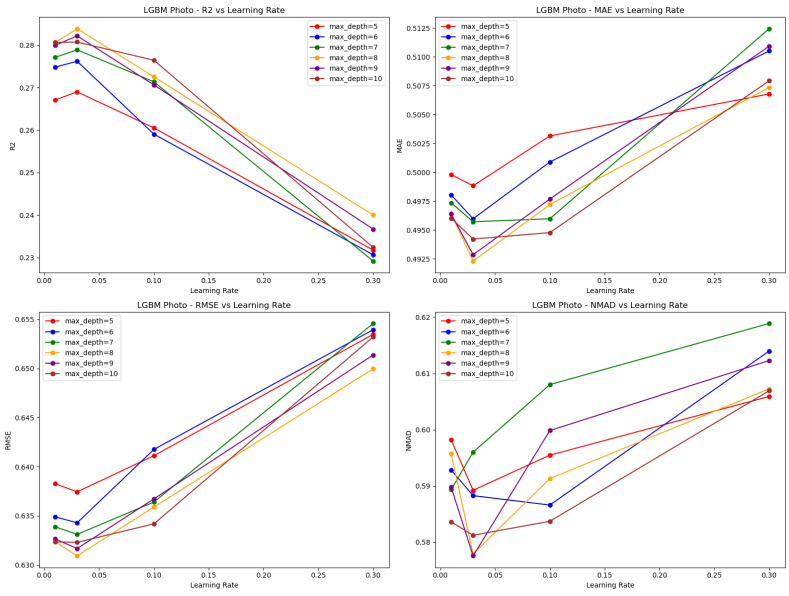


Figure 4.16 LightGBM photo: R^2, MAE, RMSE, NMAD vs. learning rate & max_depth.

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

4.7.4 Training Curves: Spectra

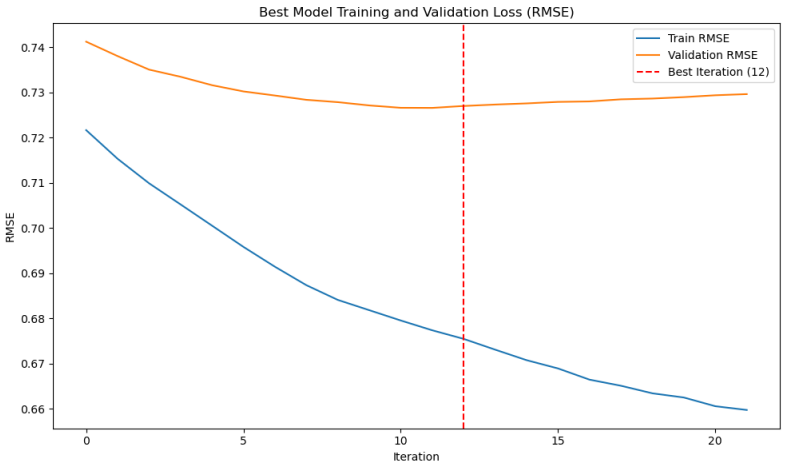


Figure 4.17 LightGBM spectra: training vs. validation RMSE; red dashed line = best iteration.

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

4.7.5 Hyperparameter Sweep: Spectra

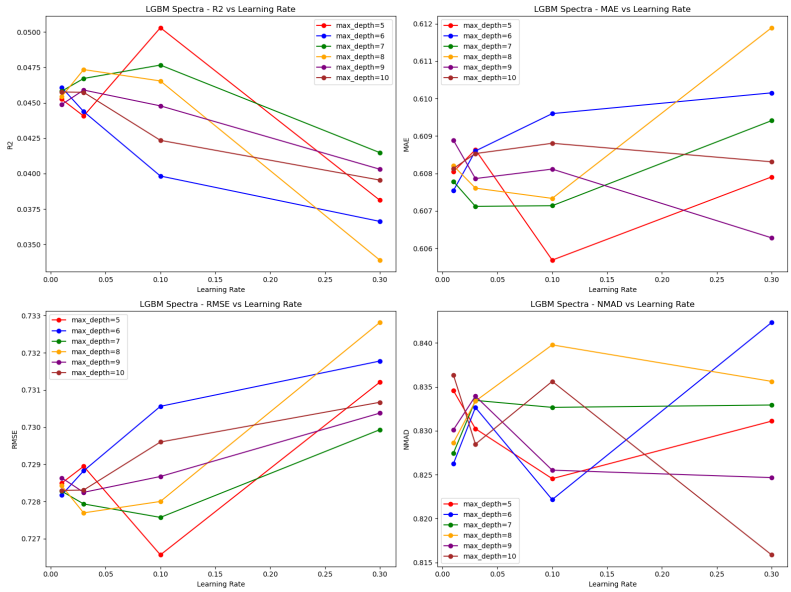


Figure 4.18 LightGBM spectra: R^2, MAE, RMSE, NMAD vs. learning rate & max_depth.

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

4.7.6 Training Curves: Early Fusion

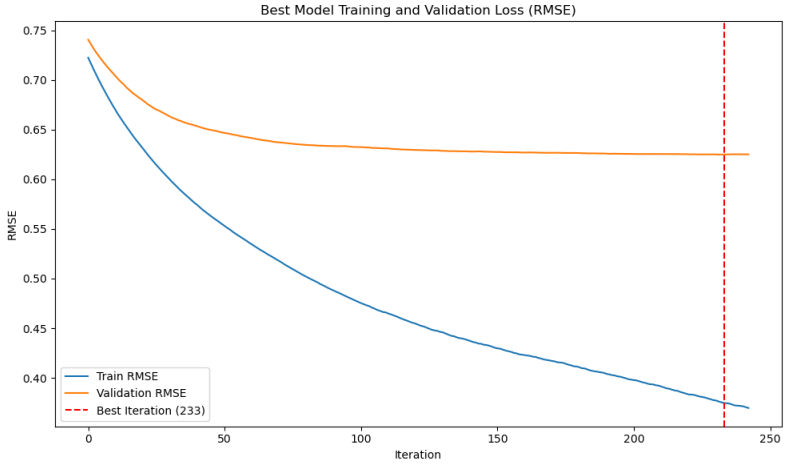


Figure 4.19 LightGBM early fusion: training vs. validation RMSE; red dashed line = best iteration.

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

4.7.7 Hyperparameter Sweep: Early Fusion

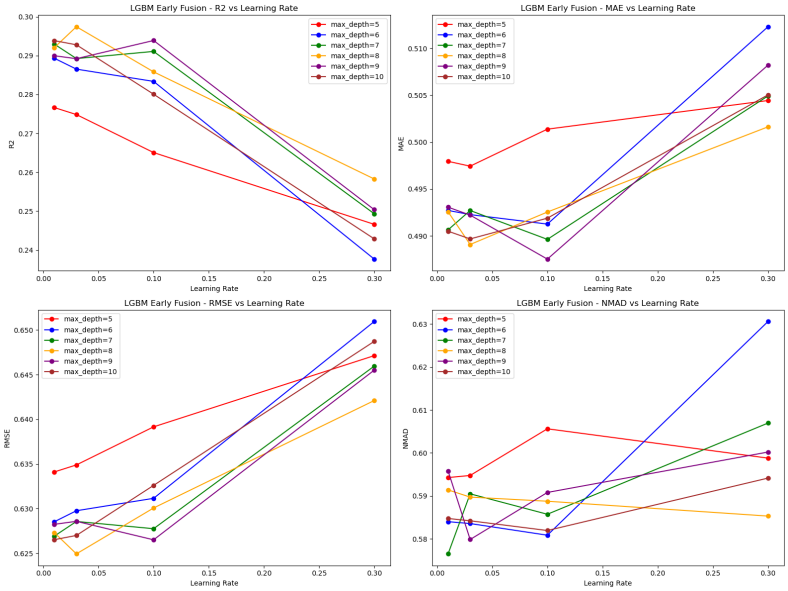


Figure 4.20 LightGBM early fusion: R^2, MAE, RMSE, NMAD vs. learning rate & max_depth.

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

4.7.8 Overall Metrics and Runtime

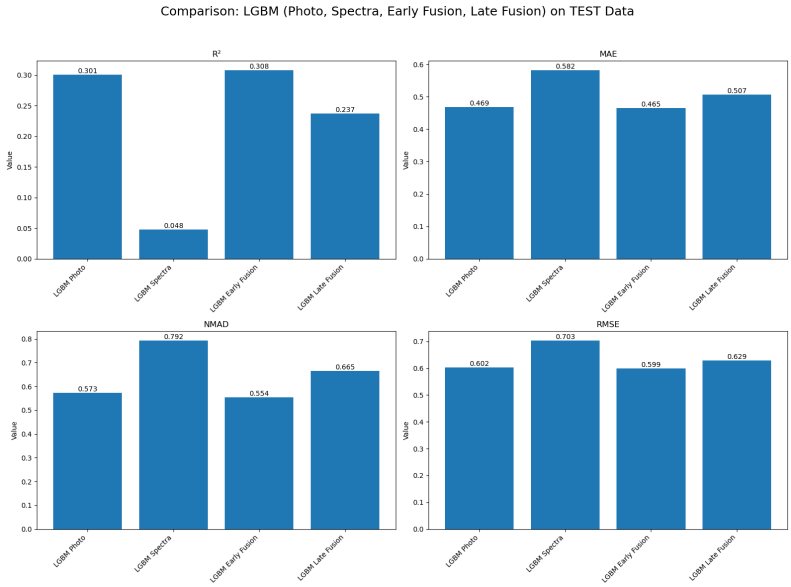


Figure 4.21 LightGBM: metric comparison across modalities.

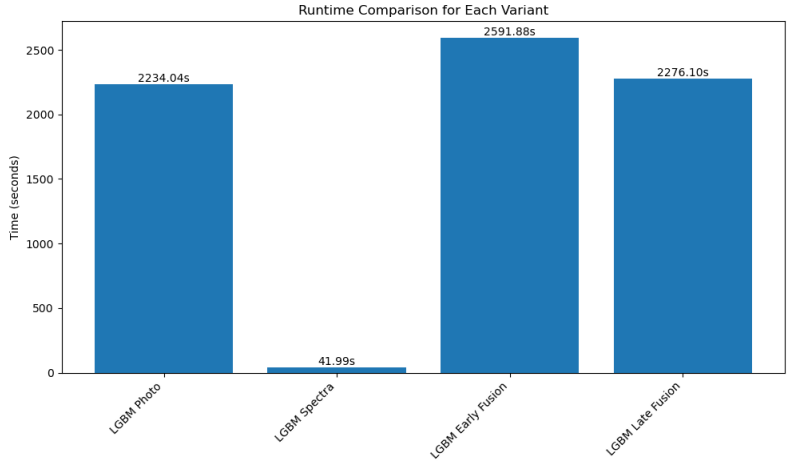


Figure 4.22 LightGBM: wall-clock runtime across modalities.

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.

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.

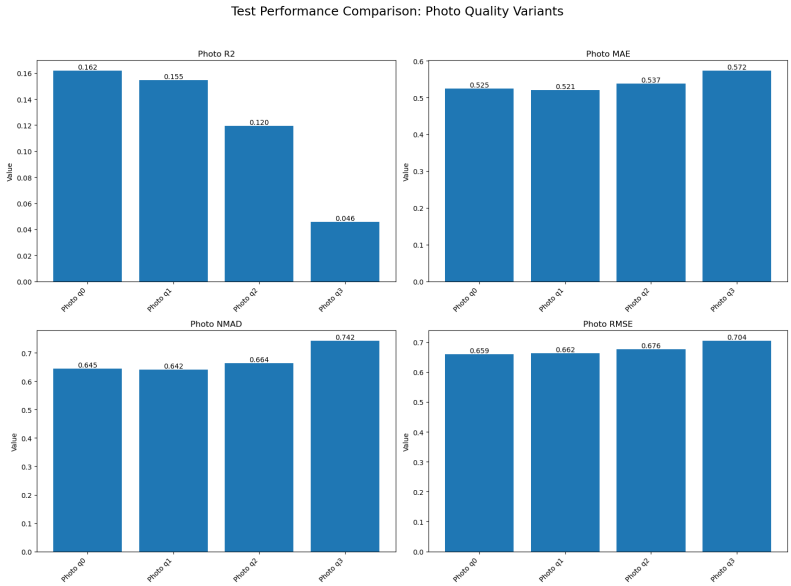


Figure 4.23 Decision Tree performance vs. image quality (q0–q3) [49].

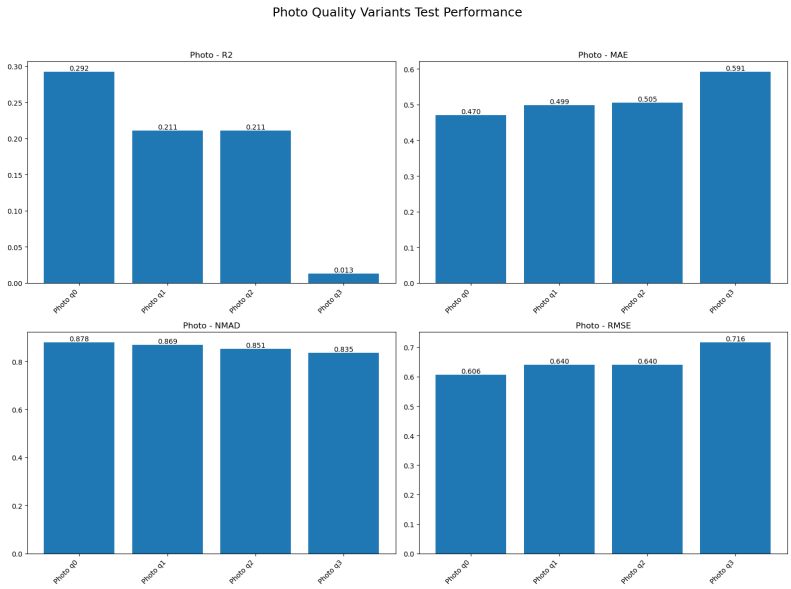


Figure 4.24 VGGNet12 performance vs. image quality (q0–q3) [50].

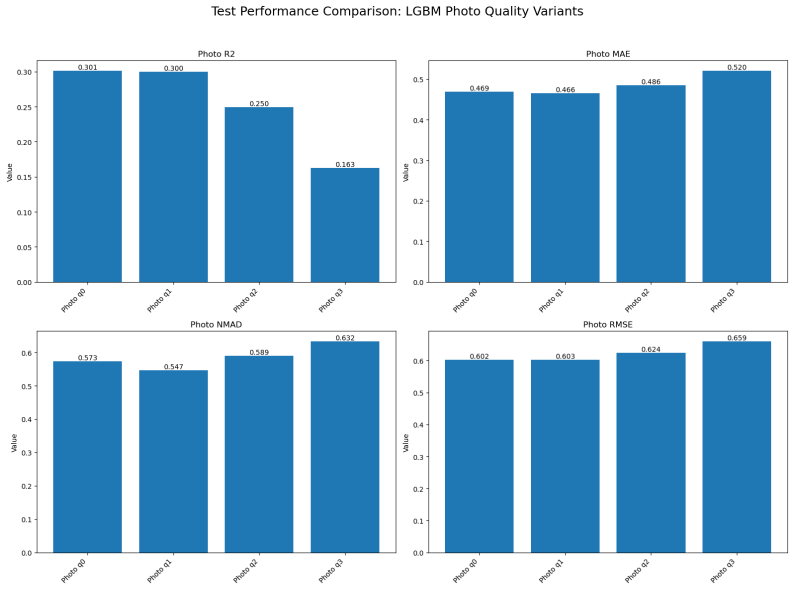


Figure 4.25 LightGBM performance vs. image quality (q0–q3) [51].

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 [52]. 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.

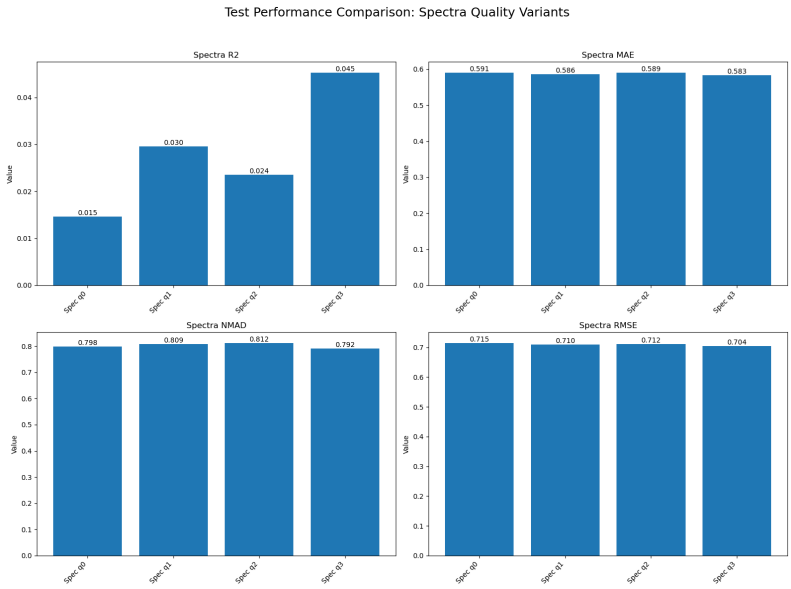


Figure 4.26 Decision Tree performance vs. spectra quality (q0–q3) [49].

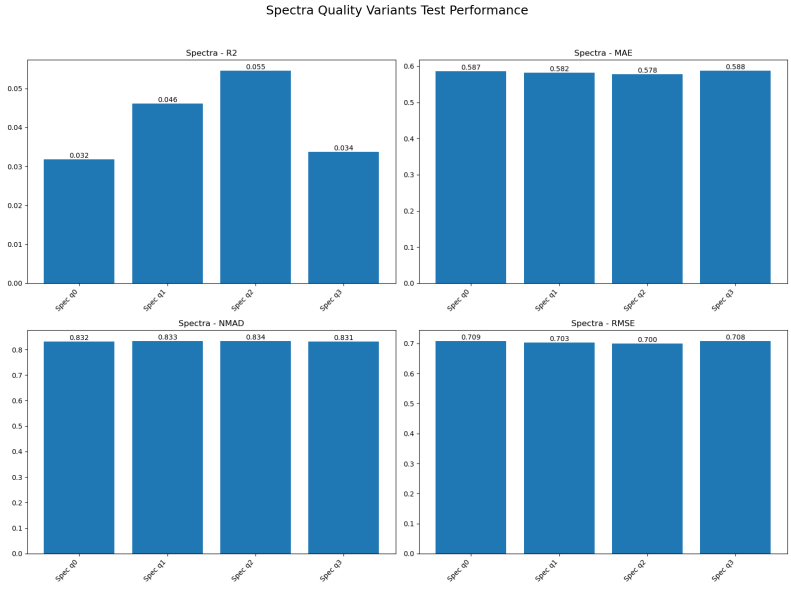


Figure 4.27 VGGNet12 performance vs. spectra quality (q0–q3) [50].

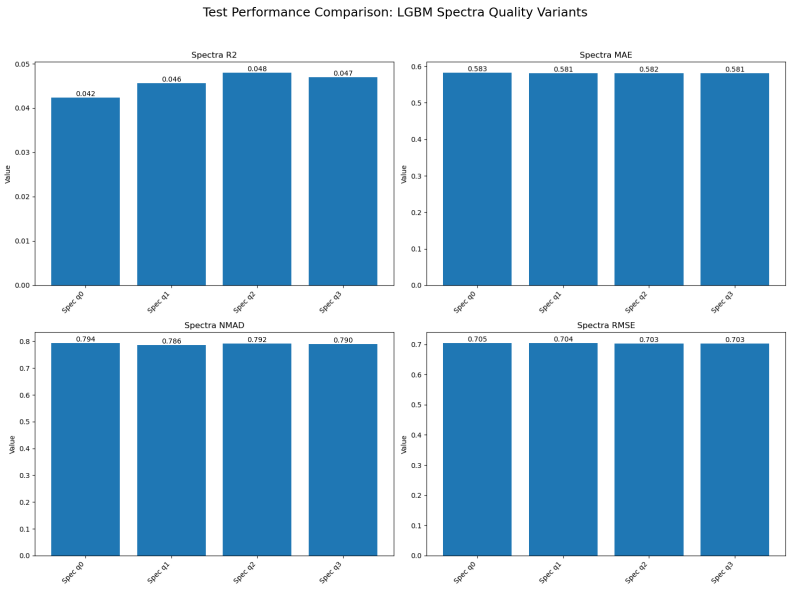


Figure 4.28 LightGBM performance vs. spectra quality (q0–q3) [51].

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

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

and late-fusion from

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

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

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.

5.1 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 cues 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. 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.
4. 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.
5. *SDSS Data Release 7* [online]. [N.d.]. Available also from: https://classic.sdss.org/dr7/. [accessed 2025-04-28].
6. NÁDVORNÍK, Jirí; ŠKODA, P; TVRDÍK, Pavel. HiSS-cube: A scalable framework for hierarchical semi-sparse cubes preserving uncertainties. *Astronomy and Computing*. 2021, vol. 36, p. 100463.
7. 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.

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

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

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

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

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

13. 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].

14. FUKUGITA, M; SHIMASAKU, K; ICHIKAWA, T; GUNN, JE, et al. *The Sloan digital sky survey photometric system*. 1996. Tech. rep. SCAN-9601313.

15. 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.

16. 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.

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

18. CALZETTI, Daniela; WU, S-Y; HONG, S; KENNICUTT, RC; LEE, JC; DALE, DA; ENGELBRACHT, CW; VAN ZEE, L; DRAINE, BT; HAO, C-N, et al. The calibration of monochromatic far-infrared star formation rate indicators. *The Astrophysical Journal*. 2010, vol. 714, no. 2, p. 1256.

19. MURPHY, EJ; CONDON, JJ; SCHINNERER, E; KENNICUTT, RC; CALZETTI, D; ARMUS, L; HELOU, G; TURNER, JL; ANIANO, G; BEIRAO, P, et al. Calibrating extinction-free star formation rate diagnostics with 33 GHz free–free emission in NGC 6946. *The Astrophysical Journal*. 2011, vol. 737, no. 2, p. 67.

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. [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. BORGES, Paulo Vinicius Koerich. *Illustration of early fusion, late fusion, and middle fusion methods... — Download Scientific Diagram*. [N.d.]. Available also from: `https://www.researchgate.net/figure/Illustration-of-early-fusion-late-fusion-and-middle-fusion-methods-used-by-multimodal_fig2_362028535`. [Online; accessed 2025-04-30].

33. 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].

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

35. 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.

36. 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.

37. 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.

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

39. 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.

40. 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.

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

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

43. 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.

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

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

46. 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.

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

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

49. *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].

50. 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].

51. *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].

52. 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.