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 22, 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.*
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.

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 22, 2025

Abstract

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

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

Contents

List of Figures

List of Tables

List of code listings

List of abbreviations

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

Chapter 1

Introduction

1.1 General Description and Relevance of the Study

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

In astrophysics, large-scale surveys like the Sloan Digital Sky Survey (SDSS) [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 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) [3]. Each image captures a specific portion of the spectrum, enabling a detailed analysis of the structural and physical properties of the objects.

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

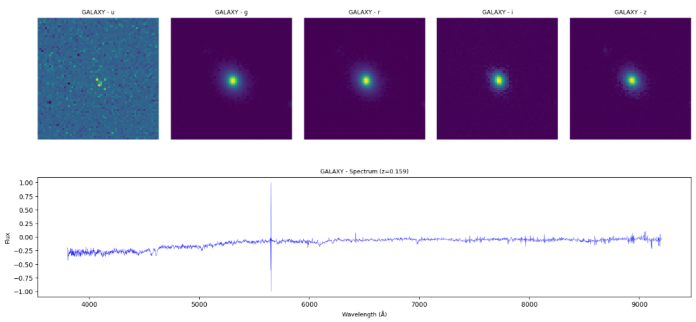

Figure 1.1 An example of an object. Top 5 pixel photos, bottom a spectrum.

1.3 Prediction of Star Formation Rate (SFR)

One of the primary objectives of this research is to predict the star formation rate (SFR) using the available SDSS data. In our dataset, the SFR is represented by the column `AVR` (mean value of the SFR distribution) [4]. By applying machine learning techniques, we aim to evaluate the feasibility of accurately predicting SFR using various types of input data.

The planned experiments include:

- Predicting SFR using only photometric images.
- Predicting SFR using only spectroscopic data.
- Employing a multimodal approach that combines both images and spectra.

1.4 Research Challenges

Working with the SDSS data presents several challenges:

1. **Data Filtering.** The original dataset contains over 4 million objects, and we need to determine which of these are suitable for machine learning.
2. **Quality of Images and Spectra.** Multiple quality levels allow optimization of the pipeline, but determining the optimal resolution is non-trivial.
3. **Multiple Objects in One Image.** Overlapping signals can degrade machine learning performance, so automatic object detection and isolation methods are needed.

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 using the Scene dataset and try to adapt findings to SDSS.

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 [5].

1.6.1.2 Why Spectral Analysis Is Needed

- **Chemical Composition:** Spectral lines reveal elemental makeup.
- **Velocity Measurements:** Line shifts indicate motion.
- **Physical Conditions:** Emission/absorption lines indicate temperature, density.

All of these diagnostics are discussed in Chapter 1 ("Why Record Spectra of Astronomical Objects?") of Tennyson's second edition [5, p. 1–6].

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

SDSS uses five broadband filters with approximate effective wavelengths of $u = 354\,\text{nm}$, $g = 477\,\text{nm}$, $r = 623\,\text{nm}$, $i = 762\,\text{nm}$ and $z = 913\,\text{nm}$ [3]

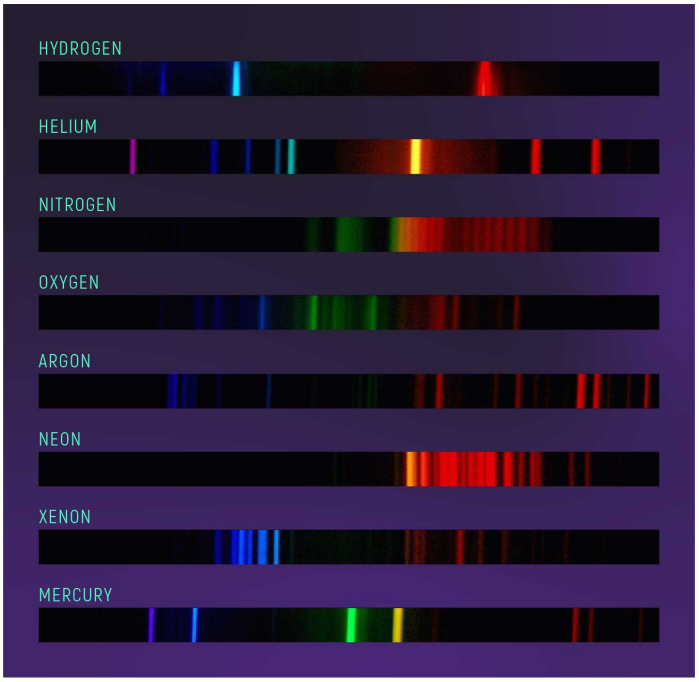

Figure 1.2 Example of atomic spectral lines for different elements.[6]

1.6.3 Star Formation Rate (SFR)

1.6.3.1 What Is SFR

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

1.6.3.2 How SFR Is Determined

Emission line luminosity, especially Hα, is used:

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

[7]

1.6.3.3 Why SFR Matters

The star formation rate governs the chemical enrichment and overall evolutionary pathway of galaxies, as well as their energy output through radiation, stellar winds, and supernova feedback (8).

1.7 Conclusion of the Introduction

In summary, this thesis explores the prediction of SFR in astronomical objects using SDSS data, leveraging both images and spectra. Multimodal preliminary studies motivate the methodology detailed in subsequent chapters.

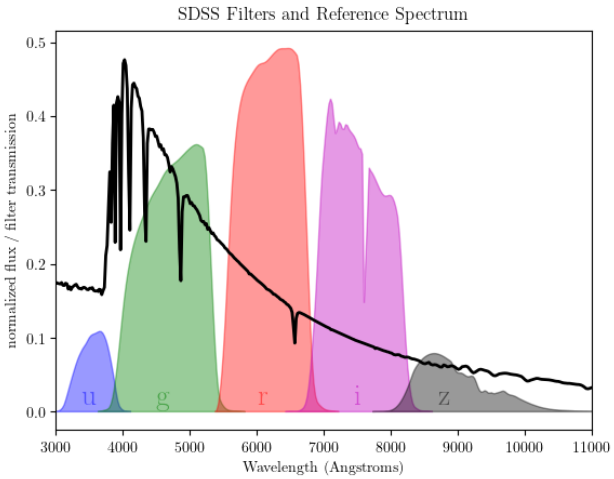

Figure 1.3 Transmission curves of the SDSS *u, g, r, i, z* filters.

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) [9] and 13,911 have `FLAG`$\neq 0$. Table 2.1 summarizes these counts [10].

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

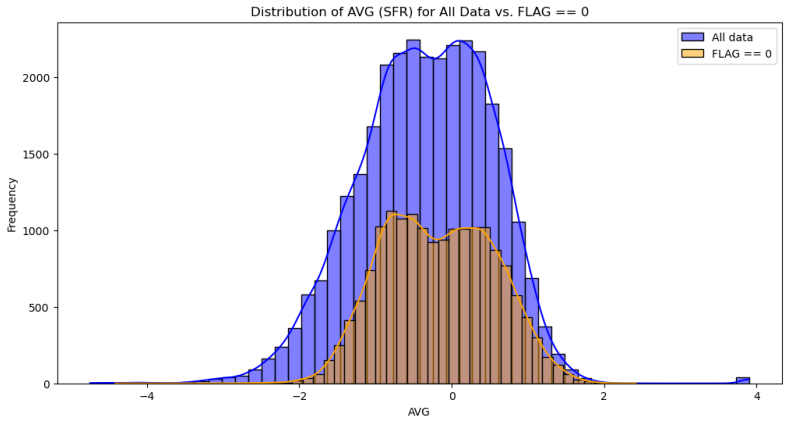

Figure 2.1 Distribution of `AVG` (log SFR) in the filtered sample [10].

2.2 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 [9] 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."
>
> We proceed exclusively with the `FLAG`=0 subset (16,841 galaxies).

2.3 Image and Spectrum Data Availability

Using the HISS-Cube [11] pipeline applied to SDSS DR7, we obtain four resolutions of imaging and spectroscopic data for each `FLAG`=0 galaxy [10] :

- Image cutouts: $(16{,}841, 5, 64, 64)$, $(16{,}841, 5, 32, 32)$, $(16{,}841, 5, 16, 16)$, $(16{,}841, 5, 8, 8)$
- Spectra: $(16{,}841, 4620)$, $(16{,}841, 2310)$, $(16{,}841, 1155)$, $(16{,}841, 577)$

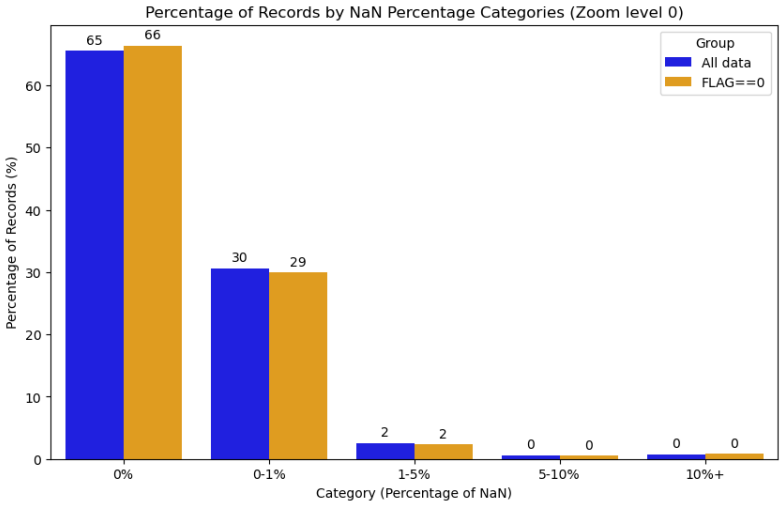

Figure 2.2 Percentage of spectra by fraction of missing (NaN) flux values at Zoom level 0 for `FLAG`=0 [10] .

2.4 Analysis of NaN Block Lengths and Positions

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

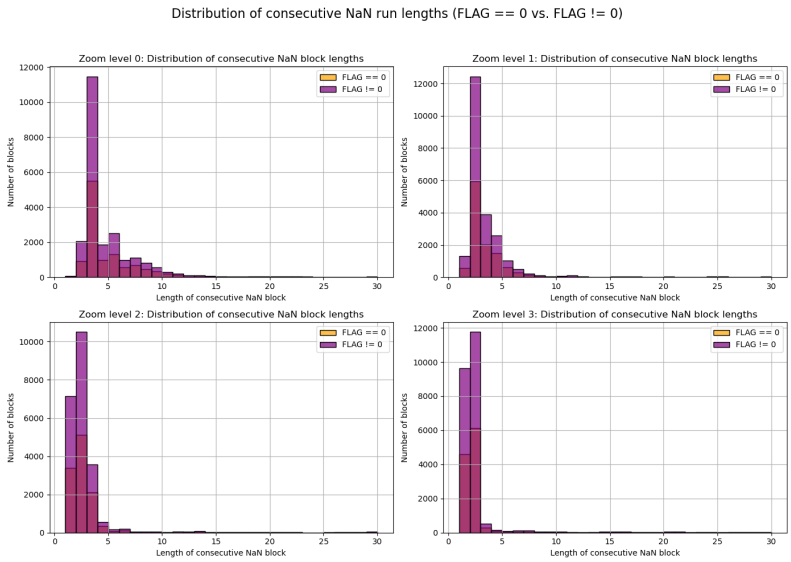

Figure 2.3 Distribution of consecutive NaN run lengths at each resolution for `FLAG`=0 [10] .

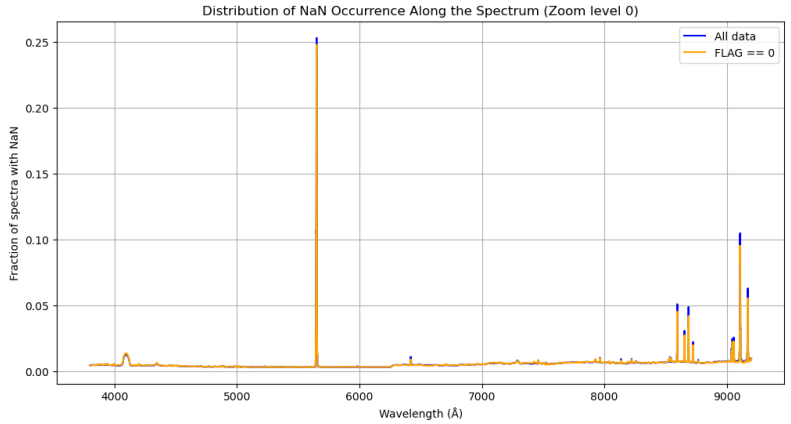

Figure 2.4 Typical wavelength regions where NaN gaps commonly occur (Zoom level 0) [10] .

2.5 Detection and Removal of Multi-Object Cutouts

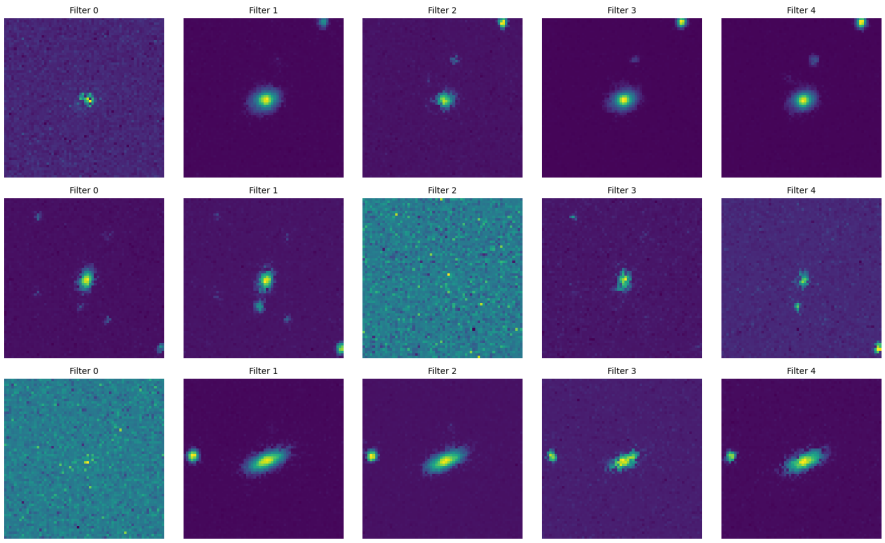

Figure 2.5 Example of a cutout containing multiple detected sources, excluded from the final sample [10] .

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 [10] 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].

2.6 Summary of Final Dataset

The cleaned dataset for supervised regression consists of:

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

- Total of 11,179 galaxies

Chapter 3

Machine Learning Methodology

3.1 Comparative Analysis: The Scene Dataset Example

To preliminarily evaluate the benefits of multimodal learning, we conducted experiments on the publicly available *Scene dataset* [14]. This dataset contains two modalities:

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

The classification task consists of two hierarchical objectives:

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

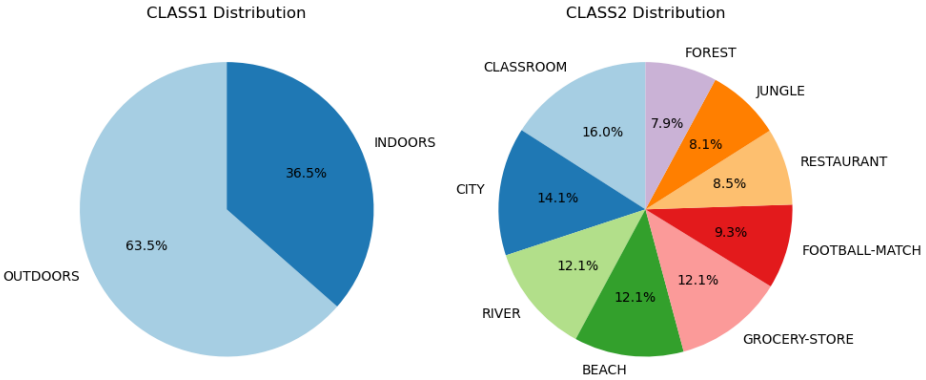

Figure 3.1 CLASS1 (left) and CLASS2 (right) label distributions for the Scene dataset [15].

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

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

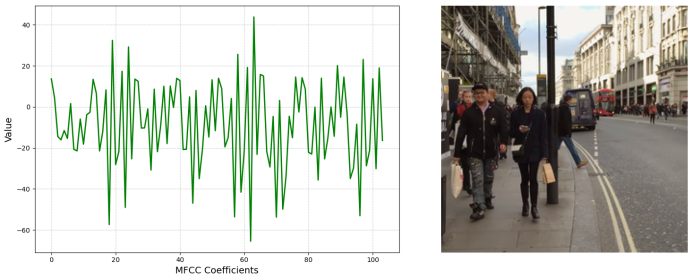

Figure 3.2 MFCC plot of the audio and street-scene photo taken during the recording [15].

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

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

3.2 Star–Galaxy–Quasar classification

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

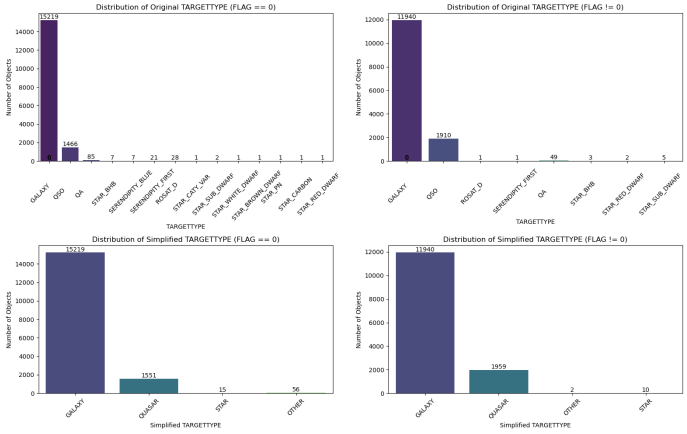

Figure 3.3 Class distribution for star–galaxy–quasar labels: galaxies outnumber quasars by a factor of 10, and stars comprise fewer than 30 objects [10].

3.3 Overview of Learning Algorithms

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

- **Decision Tree Regression (DT).** A non-parametric tree model that recursively partitions feature space by axis-aligned splits, offering interpretability and a natural baseline [16].
- **Convolutional Neural Network (VGGNet12).** A 12-layer CNN architecture that excels at large-scale image feature extraction [17].
- **Gradient Boosting Machine (LightGBM).** An efficient implementation of gradient-boosted decision trees optimized for speed and memory [18].

3.4 Experimental Setup

3.4.1 Data Splitting Strategy

We shuffle and split the cleaned sample into training, validation, and test subsets in a 60/20/20 ratio using stratified sampling on `AVG`. We then perform 5-fold cross-validation on the training set to estimate generalization error and tune hyperparameters [19, 20].

3.4.2 Preprocessing

- *Images:* pixel values are linearly scaled to $[0, 1]$ by dividing by 255 [21], then flattened for decision-tree/LightGBM models or fed as 2D arrays into VGGNet12 [22].
- *Spectra:* Any object with NaN flux values removed, yielding 11,179 gap-free spectra[23].
- *Early Fusion:* Concatenate image and spectral vectors into one feature vector [24].
- *Late Fusion:* Average photo-only and spec-only model predictions[24].

3.4.3 Hyperparameter Tuning

DT: grid search over `max_depth` $\in \{1, \ldots, 6\}$ with 5-fold CV, selecting the depth maximizing mean test R^2 [16].

VGGNet12: sweep over learning rate (`lr`) and fixed dropout=0.5, early stopping patience=30 [25, 26].

LightGBM: grid over `learning_rate` and `max_depth`, early stopping round=10 [27].

3.5 Evaluation Metrics

We evaluate all models using:

- *Coefficient of Determination (R^2).* Variance explained [16].

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

- *Mean Absolute Error (MAE).* Average absolute deviation [16].

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

- *Root Mean Square Error (RMSE).* Quadratic penalty on large errors [16].

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

- *Normalized Median Absolute Deviation (NMAD).* $1.4826 \times \text{median}(|\epsilon - \text{median}(\epsilon)|)$ [28].

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

3.6 Multimodal Fusion Strategies

3.6.1 Early Fusion

Concatenate CNN feature vector (size N_{img}) with spectral vector (size N_{spec}) into one regressor input [29].

3.6.2 Late Fusion

Average independent predictions:

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

3.7 Decision Tree Regression

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

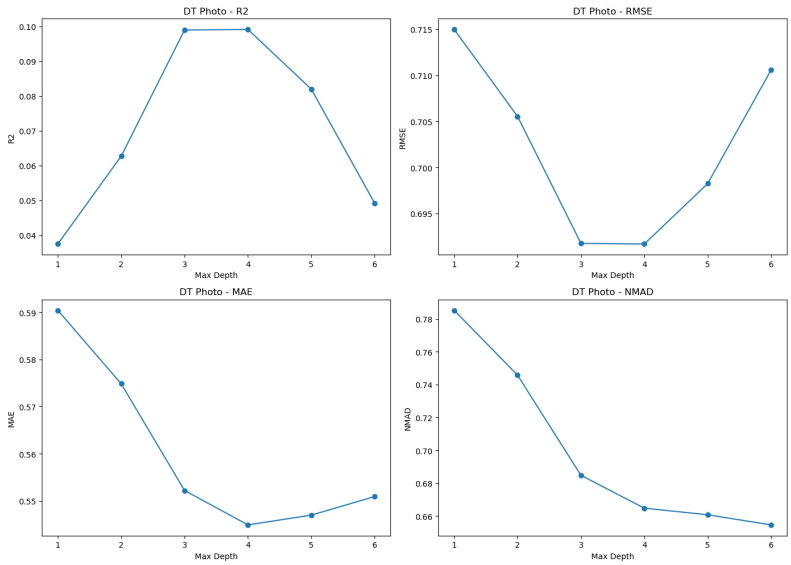

Figure 3.4 DT on photographs: R^2, MAE, RMSE, and NMAD vs. max. tree depth. Best $d = 4$ (all except NMAD).

Figure 1: Photo-only DT performance.

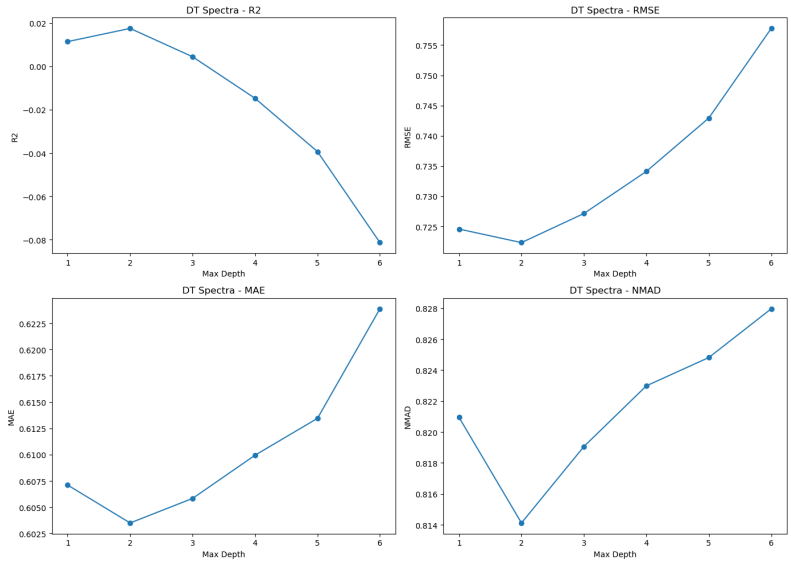

Figure 3.5 DT on spectra: R^2, MAE, RMSE, and NMAD vs. max. tree depth. Best $d = 2$.

Figure 2: Spectra-only DT performance.

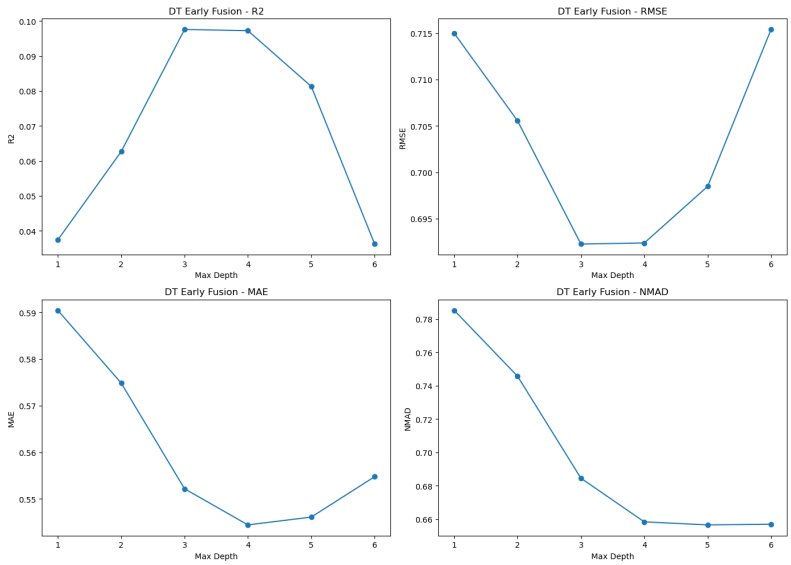

Figure 3.6 DT early fusion: R^2, MAE, RMSE, and NMAD vs. tree depth. Best $d = 3$ by R^2.

Figure 3: Early fusion DT performance.

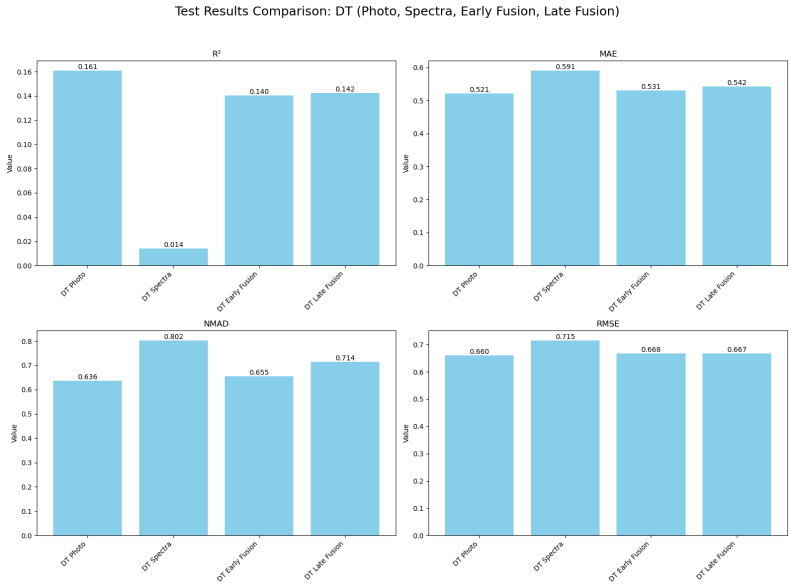

Figure 3.7 DT: metric comparison across modalities (photo, spectra, early, late).

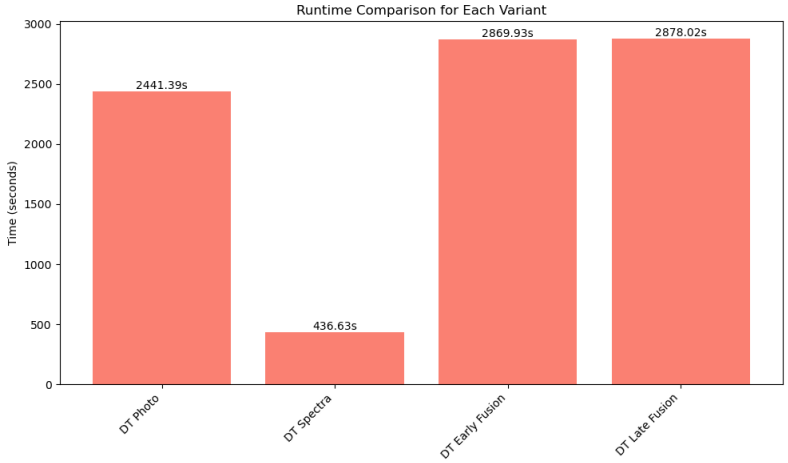

Figure 3.8 DT: wall-clock runtime across modalities.

3.8 Convolutional Neural Network: VGGNet12

The VGGNet12 model stacks 3×3 convolutions, max-pooling, then three FC layers with dropout, fine-tuned from ImageNet [17].

3.8.1 Architecture and Training Protocol

We optimize custom MSE loss,

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

using Adam, early stopping (patience=30), and focus hyperparameter tuning on learning rate [30, 26, 25].

3.8.2 Training Curves: Photographs

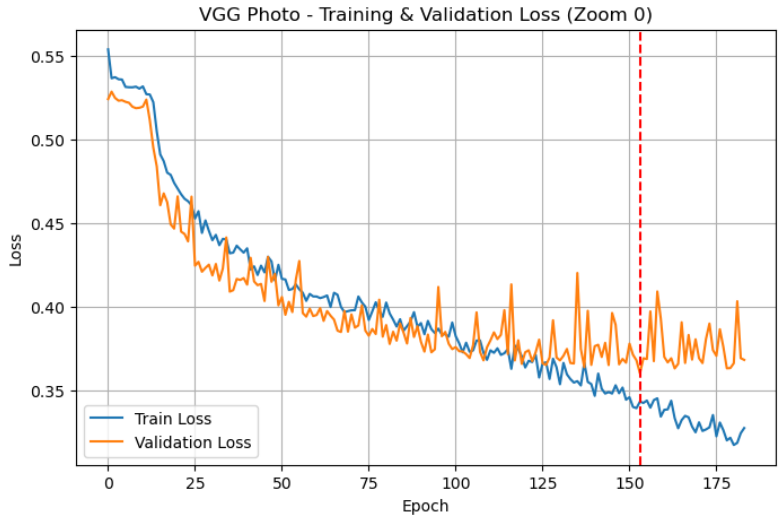

Figure 3.9 VGGNet12 photo: training (blue) vs. validation (orange) loss per epoch; red dashed line marks lowest val. loss.

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

3.8.3 Hyperparameter Sweep: Photographs

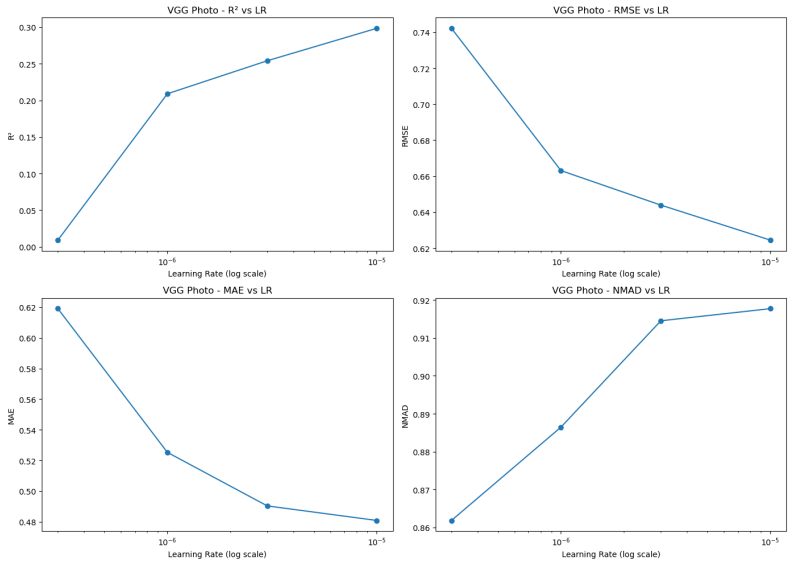

Figure 3.10 VGGNet12 photo: R^2, MAE, RMSE, NMAD vs. learning rate.

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

3.8.4 Training Curves: Spectra

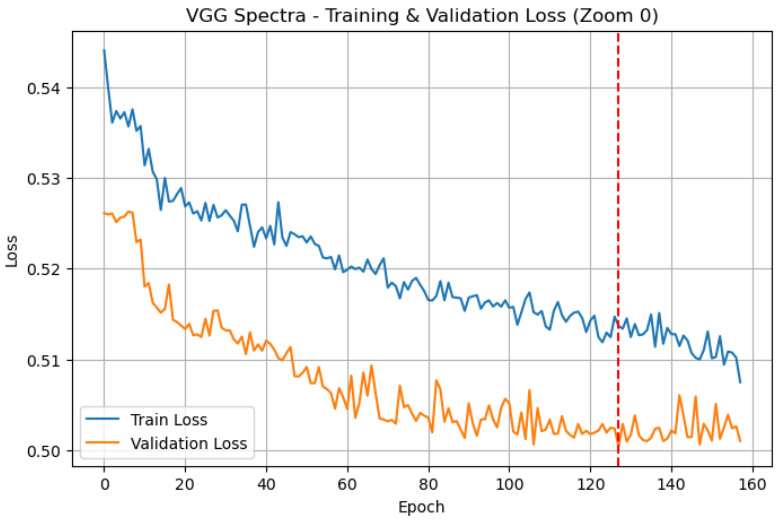

Figure 3.11 VGGNet12 spectra: training vs. validation loss per epoch; red dashed line = best epoch.

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

On this graph, it is particularly noticeable that there are epochs where the validation loss dips below the training loss. This behavior is expected in networks using dropout: during training dropout with $p = 0.5$ randomly deactivates neurons, adding noise and raising training loss, whereas no dropout is applied during validation, so the validation loss can occasionally be lower than the training loss [31].

3.8.5 Hyperparameter Sweep: Spectra

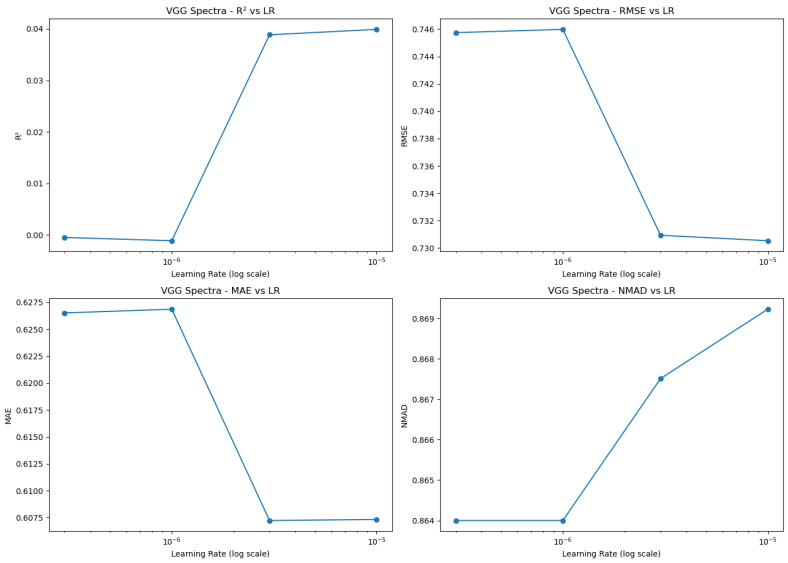

Figure 3.12 VGGNet12 spectra: R^2, MAE, RMSE, NMAD vs. learning rate.

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

3.8.6 Training Curves: Early Fusion

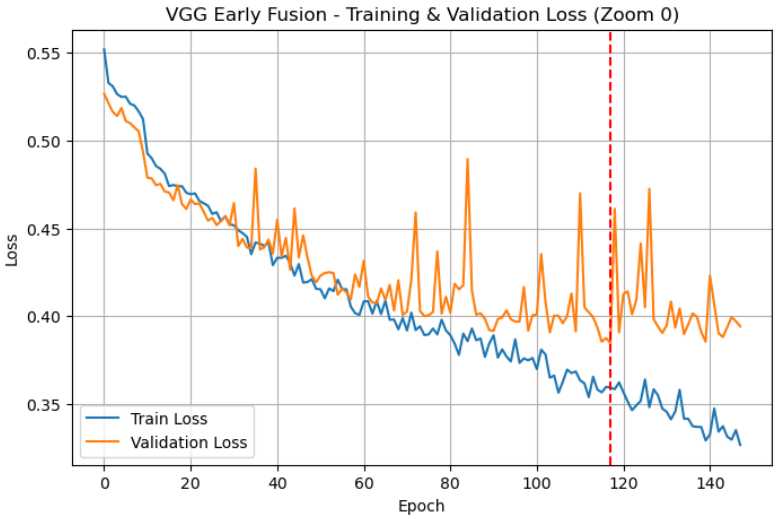

Figure 3.13 VGGNet12 early fusion: training vs. validation loss; red dashed line = best epoch.

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

3.8.7 Hyperparameter Sweep: Early Fusion

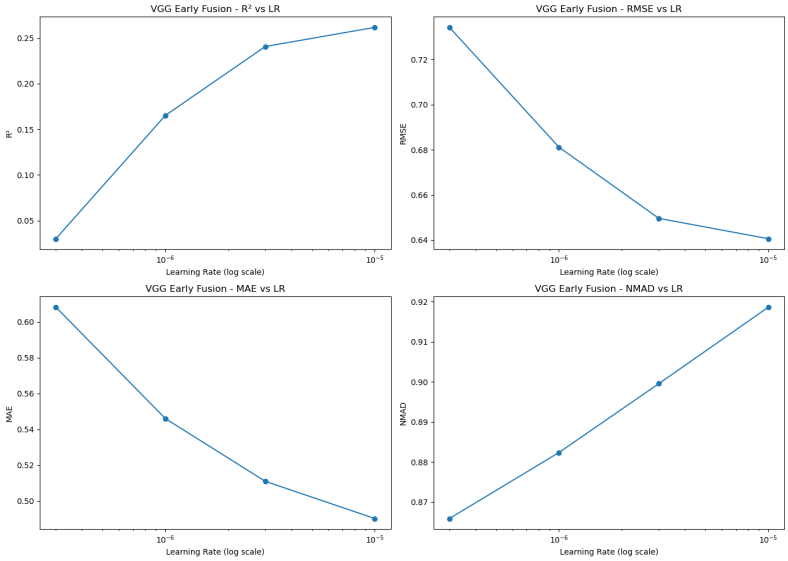

Figure 3.14 VGGNet12 early fusion: R^2, MAE, RMSE, NMAD vs. learning rate.

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

3.8.8 Overall Metrics and Runtime

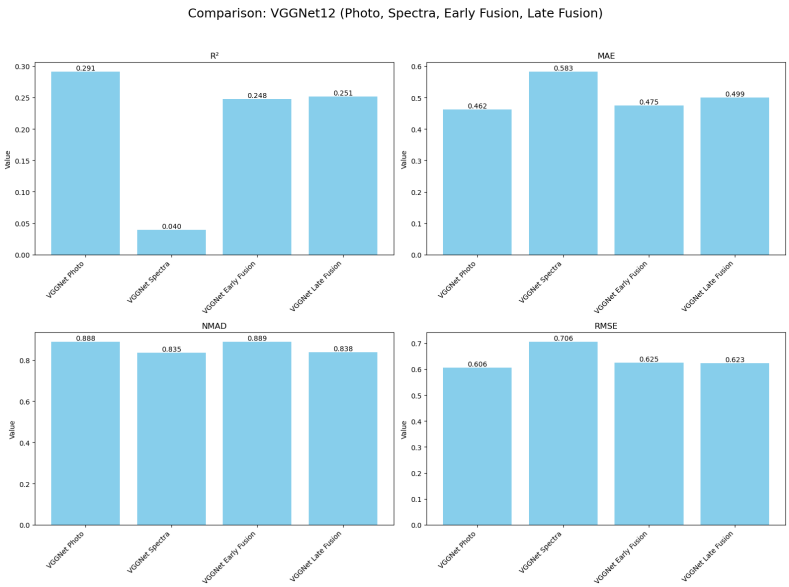

Figure 3.15 VGGNet12: metric comparison across modalities.

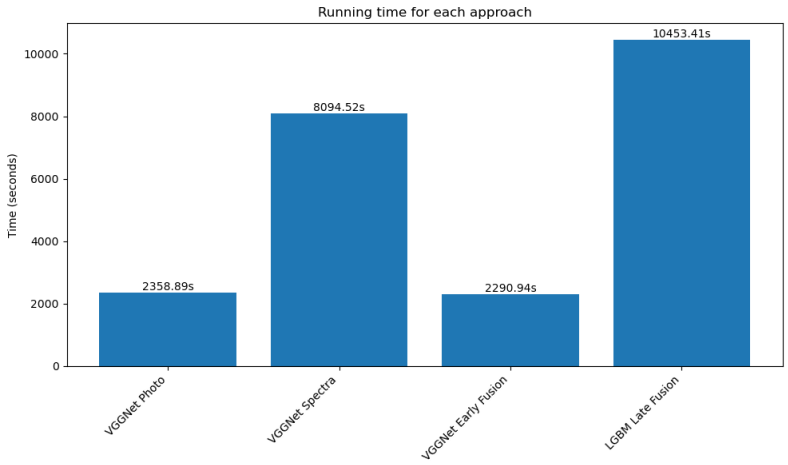

Figure 3.16 VGGNet12: wall-clock runtime across modalities.

3.9 Gradient Boosting Machine: LightGBM

LightGBM grows trees leaf-wise with histogram-based splitting and optimizes RMSE with early stopping (10 rounds) [18, 27].

3.9.1 Architecture and Training Protocol

We minimize RMSE:

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

and tune `learning_rate` and `max_depth`; early stopping prevents overfitting [**Probst2019**].

3.9.2 Training Curves: Photographs

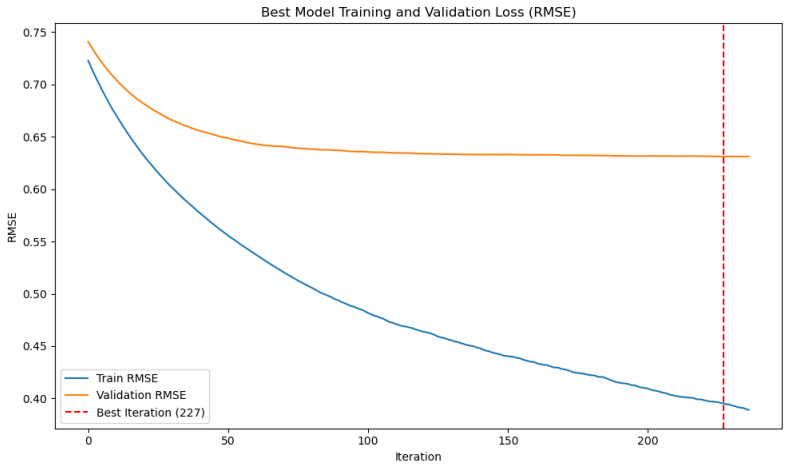

Figure 3.17 LightGBM photo: training vs. validation RMSE per iteration; red dashed line = best iteration.

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

3.9.3 Hyperparameter Sweep: Photographs

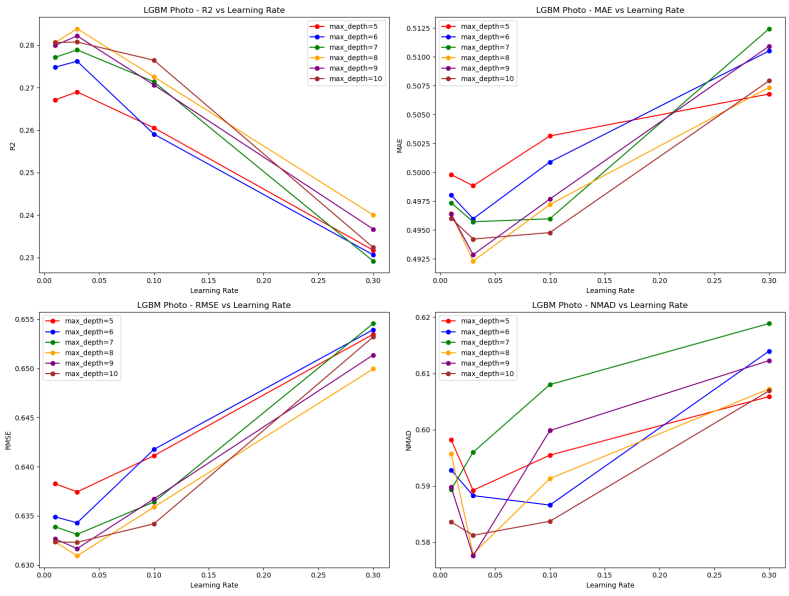

Figure 3.18 LightGBM photo: R^2, MAE, RMSE, NMAD vs. learning rate & max_depth.

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

3.9.4 Training Curves: Spectra

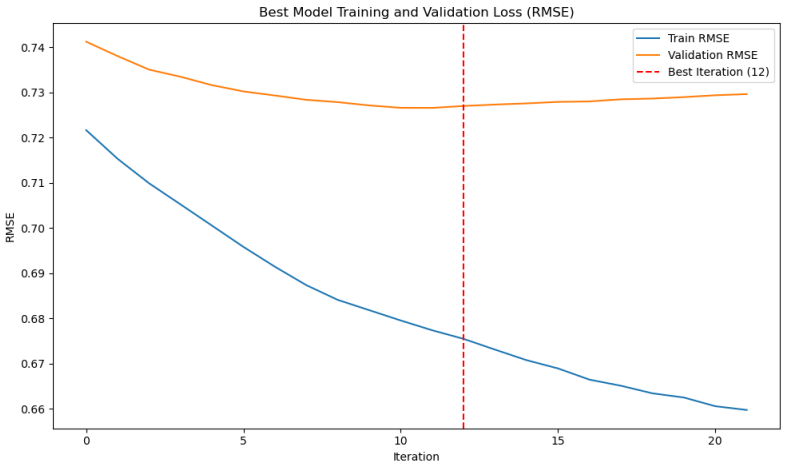

Figure 3.19 LightGBM spectra: training vs. validation RMSE; red dashed line = best iteration.

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

3.9.5 Hyperparameter Sweep: Spectra

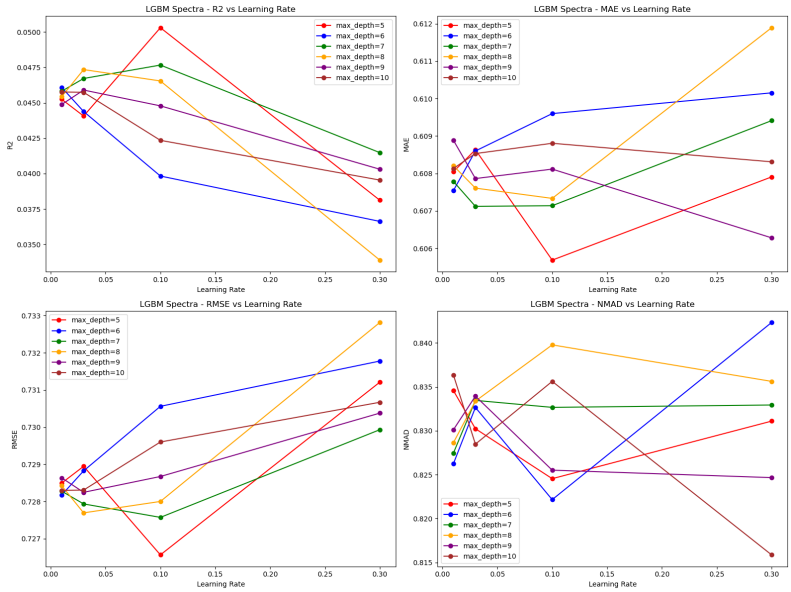

Figure 3.20 LightGBM spectra: R^2, MAE, RMSE, NMAD vs. learning rate & max_depth.

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

3.9.6 Training Curves: Early Fusion

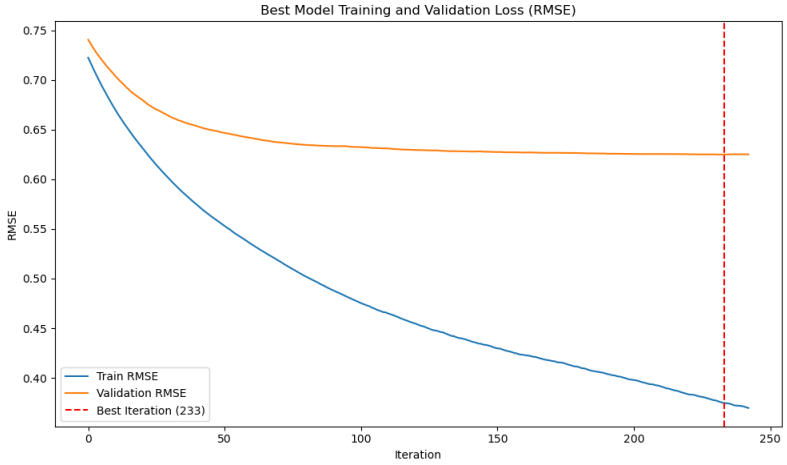

Figure 3.21 LightGBM early fusion: training vs. validation RMSE; red dashed line = best iteration.

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

3.9.7 Hyperparameter Sweep: Early Fusion

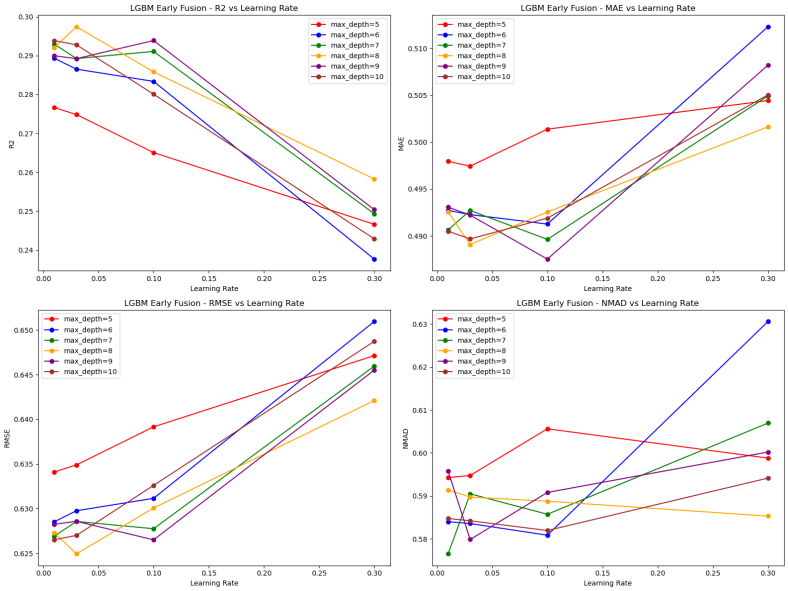

Figure 3.22 LightGBM early fusion: R^2, MAE, RMSE, NMAD vs. learning rate & max_depth.

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

3.9.8 Overall Metrics and Runtime

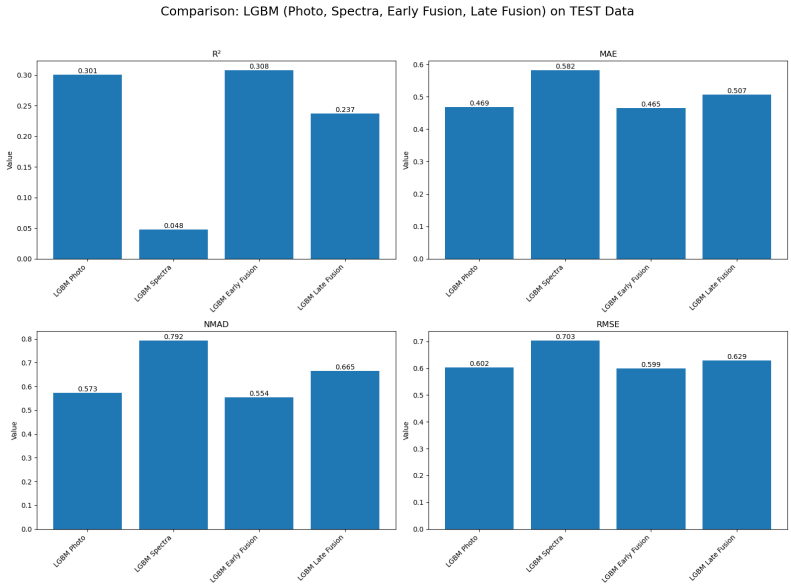

Figure 3.23 LightGBM: metric comparison across modalities.

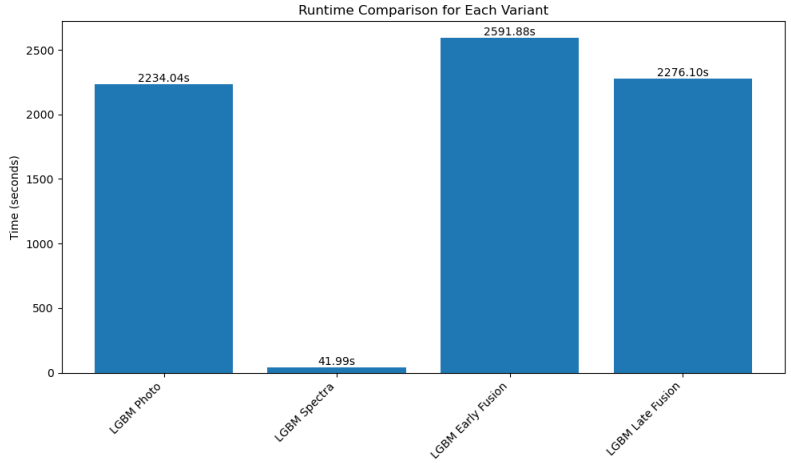

Figure 3.24 LightGBM: wall-clock runtime across modalities.

3.10 Summary and Outlook

Among all models and fusion strategies evaluated, the early-fusion LightGBM model achieved the best overall performance across metrics (highest R^2, lowest MAE and RMSE), which is consistent with the literature showing that combining complementary modalities at the feature level often yields superior predictive power [32]. Additionally, VGGNet12 applied to photometric images alone performed remarkably well, underscoring the strength of deep CNN feature extractors for morphological information in galaxy images [33].

These results demonstrate that multimodal approaches—particularly early fusion with efficient tree-based learners—can capture both spectral and visual cues essential for accurate SFR prediction. However, the complexity and diversity of astrophysical data suggest that further research is needed: exploring larger ensembles of models, advanced fusion techniques (e.g., attention-based or late-stage meta-learners), and integration of additional modalities (e.g., environmental or kinematic data) could drive even better performance.

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

Bibliography

1. 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. FUKUGITA, M; SHIMASAKU, K; ICHIKAWA, T; GUNN, JE, et al. *The Sloan digital sky survey photometric system*. 1996. Tech. rep. SCAN-9601313.
4. MPA GARCHING. *Raw data*. 2007. Available also from: `https://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/raw_data.html`. [Online; accessed 2025-04-21].
5. TENNYSON, Jonathan. *Astronomical spectroscopy: An introduction to the atomic and molecular physics of astronomical spectroscopy*. World Scientific, 2019.
6. INSTITUTE, Space Telescope Science. *Spectroscopy 101 – Types of Spectra and Spectroscopy — Webb*. 2022. Available also from: `https://webbtelescope.org/contents/articles/spectroscopy-101--types-of-spectra-and-spectroscopy?page=1&keyword=Stars`. [Online; accessed 2025-04-22].
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. SCIKIT-LEARN DEVELOPERS. *2.3.4. Classification: Learning Labels of Astronomical Sources — scikit-learn 0.11-git documentation*. 2011. Available also from: `https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/tutorial/astronomy/classification.html`. [Online; accessed 2025-04-22].
9. MPA GARCHING. SDSS DR7 SFR documentation. *MPA Garching Web Resource*. 2007. Available also from: `https://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/sfrs.html`.
10. RUSTAMOV, Farukh. *Jupyter Notebook: ≪data$_e$xploring ≫* [`https://gitlab.fit.cvut.cz/rustafar/astronomical_data_ml/-/blob/main/SDSS/data_exploring.ipynb`]. 2025. [Online; accessed 2025-04-22].
11. NÁDVORNÍK, Jirí; ŠKODA, Petr; TVRDÍK, Pavel. HDF5 Parallelization for Hierarchical Semi-Sparse Data Cubes. In: *Astronomical Society of the Pacific Conference Series*. 2024, vol. 535, p. 115.
12. 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.
13. GONZALEZ, Rafael C. *Digital image processing*. Pearson education india, 2009.
14. BIRD, Jordan J. *Scene Classification: Images and Audio*. 2020. Available also from: `https://www.kaggle.com/datasets/birdy654/scene-classification-images-and-audio/`. [Online; accessed 2025-04-21].
15. RUSTAMOV, Farukh. *Jupyter Notebook: ≪scene≫* [`https://gitlab.fit.cvut.cz/rustafar/astronomical_data_ml/-/blob/main/Scene/scene.ipynb`]. 2025. [Online; accessed 2025-04-22].
16. HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome H; FRIEDMAN, Jerome H. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.
17. SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 2014.
18. 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.
19. 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.

20. 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.

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

22. 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.

23. 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.

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

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

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

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

28. 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.

29. 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.

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

31. 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.

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

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