Bachelor’s thesis

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

**Bc. Farukh Rustamov**

Faculty of Information Technology
Department of Theoretical Computer Science
Supervisor: doc. Ing. Damien Zlo, Ph.D.
April 21, 2025

Replace the contents of this file with official assignment.
Místo tohoto souboru sem patří list se zadáním závěrečné práce.

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis: Rustamov Farukh. *Application of Machine Learning to Predict Star Formation Rates in SDSS Data.* Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2025.

*I would like to thank my supervisor for valuable guidance and support throughout this work.*

# Declaration

I hereby declare that this thesis is my own original work and that I have properly cited all sources.

In Prague on April 21, 2025

## Abstract

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

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

## Abstrakt

**Klíčová slova**

# Contents

**1 Introduction** 1
1.1 General Description and Relevance of the Study 1
1.2 SDSS Data Description 1
1.3 Prediction of Star Formation Rate (SFR) 2
1.4 Research Challenges 2
1.5 Objectives and Tasks 3
1.6 Terminology and Illustrations 3
1.6.1 Spectra and Spectral Analysis 3
1.6.1.1 Definition of a Spectrum 3
1.6.1.2 Why Spectral Analysis Is Needed 4
1.6.2 The SDSS *u, g, r, i, z* Filters 4
1.6.3 Star Formation Rate (SFR) 4
1.6.3.1 What Is SFR 4
1.6.3.2 How SFR Is Determined 5
1.6.3.3 Why SFR Matters 5
1.7 Conclusion of the Introduction 5
1.8 Outline of Future Work and Machine Learning Approach 6
1.8.1 Data Acquisition and Preparation 6
1.8.2 HiSS-Cube Implementation 6
1.8.3 Baseline Models 6
1.8.4 Fusion Strategies 6
1.8.5 Regularization 6
1.8.6 Experimental Validation 6

**2 Data Exploration** 7
2.1 Dataset Overview and Initial Filtering 7
2.2 SFR Estimation Quality: `FLAG` Keyword 7
2.3 Image and Spectrum Data Availability 8
2.4 Analysis of NaN Block Lengths and Positions 8
2.5 Detection and Removal of Multi-Object Cutouts 11
2.6 Summary of Final Dataset 11

**3 Machine Learning Methodology** 12
3.1 Comparative Analysis: The Scene Dataset Example 12
3.2 Overview of Learning Algorithms 14
3.3 Experimental Setup 14

3.3.1 Data Splitting Strategy . . . . . . . . . . . . . . . . . . . 14
3.3.2 Preprocessing . . . . . . . . . . . . . . . . . . . . . . . . . . 14
3.3.3 Hyperparameter Tuning . . . . . . . . . . . . . . . . . . . 14
3.4 Evaluation Metrics . . . . . . . . . . . . . . . . . . . . . . . . . 15
3.5 Multimodal Fusion Strategies . . . . . . . . . . . . . . . . . . . 15
3.5.1 Early Fusion . . . . . . . . . . . . . . . . . . . . . . . . . 15
3.5.2 Late Fusion . . . . . . . . . . . . . . . . . . . . . . . . . . 15
3.6 Decision Tree Regression . . . . . . . . . . . . . . . . . . . . . . 15
3.7 Convolutional Neural Network: VGGNet12 . . . . . . . . . . . 15
3.7.1 Architecture and Training Protocol . . . . . . . . . . . . 18
3.7.2 Training Curves: Photographs . . . . . . . . . . . . . . . 18
3.7.3 Hyperparameter Sweep: Photographs . . . . . . . . . . . 18
3.7.4 Training Curves: Spectra . . . . . . . . . . . . . . . . . . 18
3.7.5 Hyperparameter Sweep: Spectra . . . . . . . . . . . . . . 18
3.7.6 Training Curves: Early Fusion . . . . . . . . . . . . . . . 21
3.7.7 Hyperparameter Sweep: Early Fusion . . . . . . . . . . . 21
3.7.8 Overall Metrics and Runtime . . . . . . . . . . . . . . . . 21
3.8 Gradient Boosting Machine: LightGBM . . . . . . . . . . . . . 21
3.8.1 Architecture and Training Protocol . . . . . . . . . . . . 21
3.8.2 Training Curves: Photographs . . . . . . . . . . . . . . . 23
3.8.3 Hyperparameter Sweep: Photographs . . . . . . . . . . . 23
3.8.4 Training Curves: Spectra . . . . . . . . . . . . . . . . . . 24
3.8.5 Hyperparameter Sweep: Spectra . . . . . . . . . . . . . . 24
3.8.6 Training Curves: Early Fusion . . . . . . . . . . . . . . . 24
3.8.7 Hyperparameter Sweep: Early Fusion . . . . . . . . . . . 24
3.8.8 Overall Metrics and Runtime . . . . . . . . . . . . . . . . 24
3.9 Summary and Outlook . . . . . . . . . . . . . . . . . . . . . . . 28

# List of Figures

1.1 An example of an object. Top 5 pixel photos, bottom a spectrum. 2
1.2 White-light decomposition through a prism. . . . . . . . . . . . 3
1.3 Example of atomic spectral lines for different elements. . . . . . 4
1.4 Transmission curves of the SDSS *u, g, r, i, z* filters. . . . . . 5

2.1 Distribution of `AVG` (log SFR) in the filtered sample. . . . . . . 8
2.2 Percentage of spectra by fraction of missing (NaN) flux values at Zoom level 0 for `FLAG`=0. . . . . . . . . . . . . . . . . . . . . . 9
2.3 Distribution of consecutive NaN run lengths at each resolution for `FLAG`=0. . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
2.4 Typical wavelength regions where NaN gaps commonly occur (Zoom level 0). . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
2.5 Example of a cutout containing multiple detected sources, excluded from the final sample. . . . . . . . . . . . . . . . . . . . . 10

3.1 CLASS1 (left) and CLASS2 (right) label distributions for the Scene dataset. . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
3.2 MFCC plot of the audio and street-scene photo taken during the recording . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
3.3 DT on photographs: $R^2$, MAE, RMSE, and NMAD vs. max. tree depth. Best $d = 4$ (all except NMAD). . . . . . . . . . . . 16
3.4 DT on spectra: $R^2$, MAE, RMSE, and NMAD vs. max. tree depth. Best $d = 2$. . . . . . . . . . . . . . . . . . . . . . . . . . 16
3.5 DT early fusion: $R^2$, MAE, RMSE, and NMAD vs. tree depth. Best $d = 3$ by $R^2$. . . . . . . . . . . . . . . . . . . . . . . . . 17
3.6 DT: metric comparison across modalities (photo, spectra, early, late). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17
3.7 DT: wall-clock runtime across modalities. . . . . . . . . . . . . 18
3.8 VGGNet12 photo: training (blue) vs. validation (orange) loss per epoch; red dashed line marks lowest val. loss. . . . . . . . . 19
3.9 VGGNet12 photo: $R^2$, MAE, RMSE, NMAD vs. learning rate. 19
3.10 VGGNet12 spectra: training vs. validation loss per epoch; red dashed line = best epoch. . . . . . . . . . . . . . . . . . . . . . 20
3.11 VGGNet12 spectra: $R^2$, MAE, RMSE, NMAD vs. learning rate. 20
3.12 VGGNet12 early fusion: training vs. validation loss; red dashed line = best epoch. . . . . . . . . . . . . . . . . . . . . . . . . 21

3.13 VGGNet12 early fusion: $R^2$, MAE, RMSE, NMAD vs. learning rate. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22
3.14 VGGNet12: metric comparison across modalities. . . . . . . . . 22
3.15 VGGNet12: wall-clock runtime across modalities. . . . . . . . . 23
3.16 LightGBM photo: training vs. validation RMSE per iteration; red dashed line = best iteration. . . . . . . . . . . . . . . . . . . 23
3.17 LightGBM photo: $R^2$, MAE, RMSE, NMAD vs. learning rate & max_depth. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24
3.18 LightGBM spectra: training vs. validation RMSE; red dashed line = best iteration. . . . . . . . . . . . . . . . . . . . . . . . . . 25
3.19 LightGBM spectra: $R^2$, MAE, RMSE, NMAD vs. learning rate & max_depth. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25
3.20 LightGBM early fusion: training vs. validation RMSE; red dashed line = best iteration. . . . . . . . . . . . . . . . . . . . . . . . . . 26
3.21 LightGBM early fusion: $R^2$, MAE, RMSE, NMAD vs. learning rate & max_depth. . . . . . . . . . . . . . . . . . . . . . . . . . 26
3.22 LightGBM: metric comparison across modalities. . . . . . . . . 27
3.23 LightGBM: wall-clock runtime across modalities. . . . . . . . . 27

# List of Tables

2.1 Record counts at successive filtering stages. . . . . . . . . . . . 7
2.2 NaN block statistics for FLAG=0 at each zoom level. . . . . . . 8

# 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) 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)** [1] in galaxies using data from SDSS [2]. 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 $z$). 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.

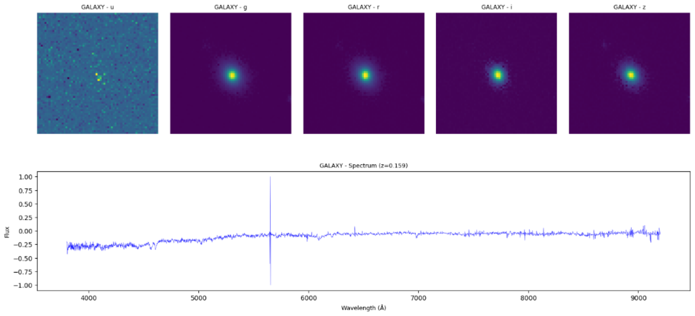

**Figure 1.1** An example of an object. Top 5 pixel photos, bottom a spectrum.

## 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) [3]. 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 comprises over 150,000 objects, but to ensure the reliability of the SFR estimates, only objects with `FLAG` $= 0$ are used.
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 using the `FLAG` parameter.
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 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.

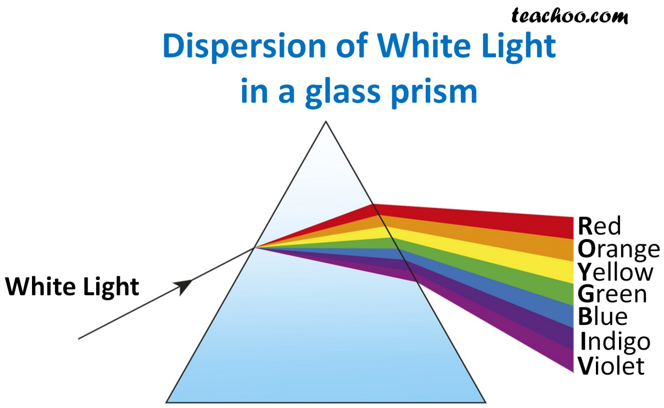

**Figure 1.2** White-light decomposition through a prism.

#### 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.

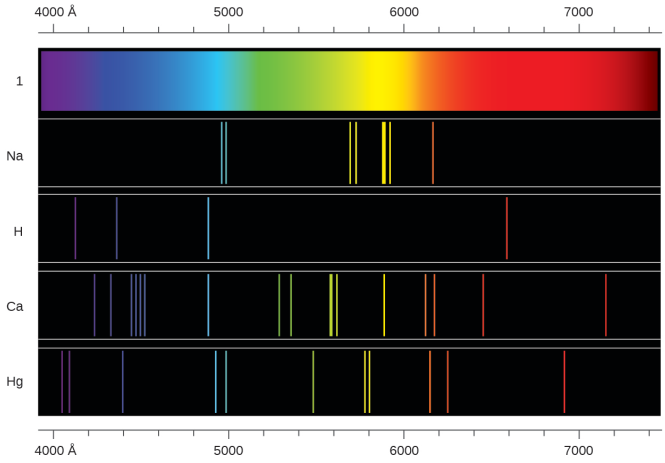

**Figure 1.3** Example of atomic spectral lines for different elements.

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

SDSS uses five broadband filters:

- **u**: 354 nm
- **g**: 477 nm
- **r**: 623 nm
- **i**: 762 nm
- **z**: 913 nm

### 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 ($M_\odot\,\mathrm{yr}^{-1}$).

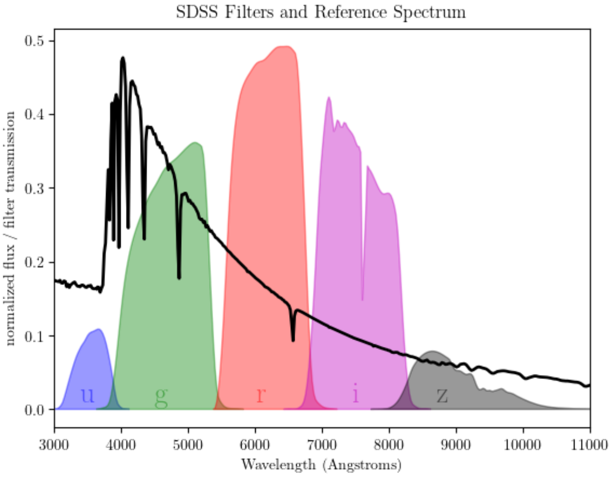

**Figure 1.4** Transmission curves of the SDSS $u$, $g$, $r$, $i$, $z$ filters.

### 1.6.3.2 How SFR Is Determined

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

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

### 1.6.3.3 Why SFR Matters

- **Galactic Evolution**
- **Energy Output**

## 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.

## 1.8 Outline of Future Work and Machine Learning Approach

### 1.8.1 Data Acquisition and Preparation

- Data exploration and visualization.
- Addressing multi-object cutouts.
- Statistical analysis of `AVR`.

### 1.8.2 HiSS-Cube Implementation

- Hierarchical HDF5 data structuring.
- Metadata association.

### 1.8.3 Baseline Models

- Spectral decision tree and hybrid network.
- Five-band CNN.

### 1.8.4 Fusion Strategies

- Early Fusion: PCA + FC layers.
- Intermediate Fusion: separate CNN and MLP branches.
- Late Fusion: averaging independent models.

### 1.8.5 Regularization

Dropout, early stopping, PCA preprocessing.

### 1.8.6 Experimental Validation

Controlled experiments, RMSE, MAE, $R^2$.

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`≠0. Table 2.1 summarizes these counts.

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

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

## 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 and for statistical studies in particular, it

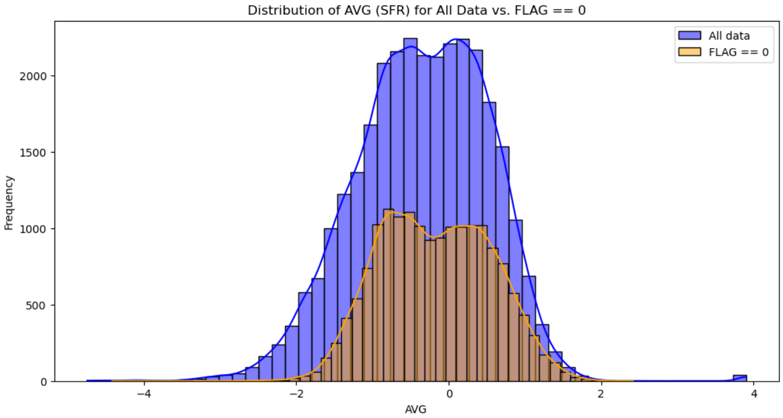

**Figure 2.1** Distribution of `AVG` (log SFR) in the filtered sample.

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.” [4]

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

## 2.3 Image and Spectrum Data Availability

Using the HISS-Cube [5] pipeline applied to SDSS DR7, we obtain four resolutions of imaging and spectroscopic data for each `FLAG`=0 galaxy:

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

## 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 |

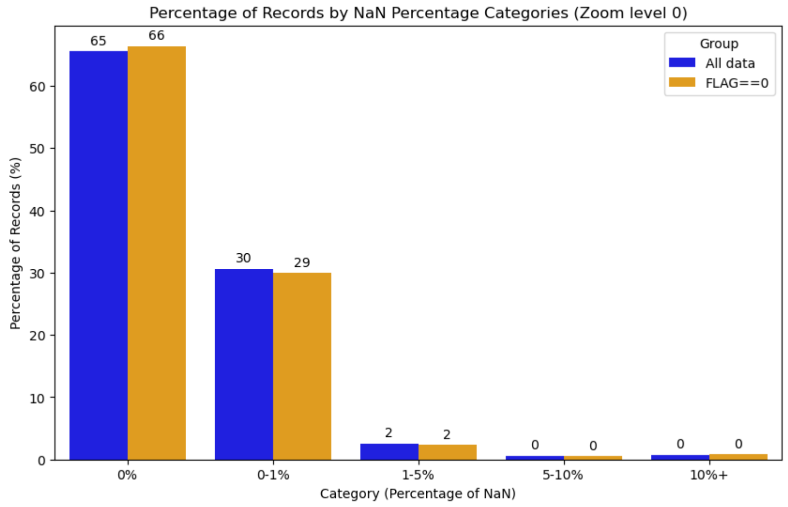

**Figure 2.2** Percentage of spectra by fraction of missing (NaN) flux values at Zoom level 0 for `FLAG`=0.

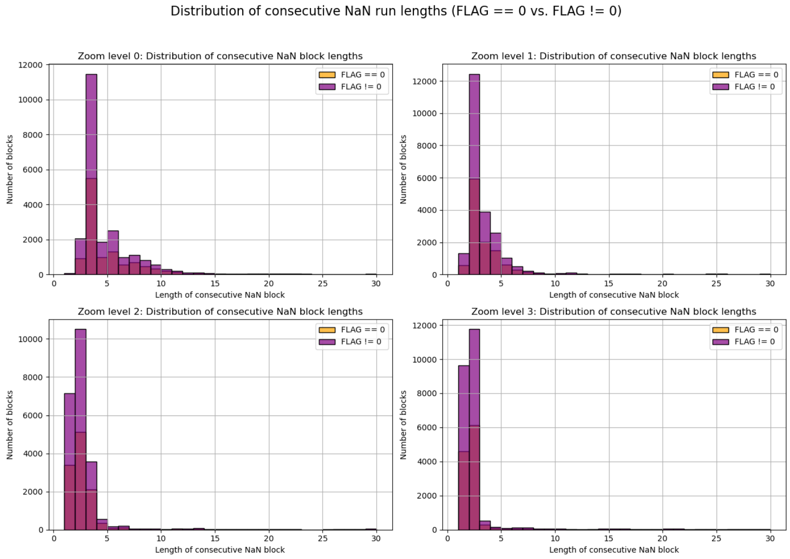

**Figure 2.3** Distribution of consecutive NaN run lengths at each resolution for `FLAG`=0.

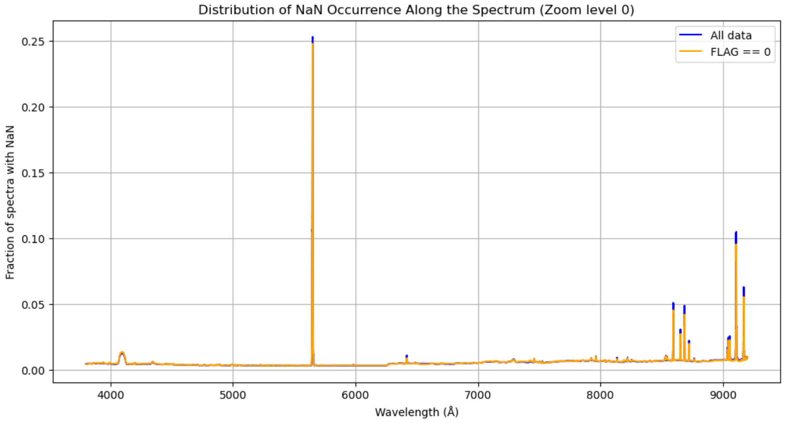

**Figure 2.4** Typical wavelength regions where NaN gaps commonly occur (Zoom level 0).

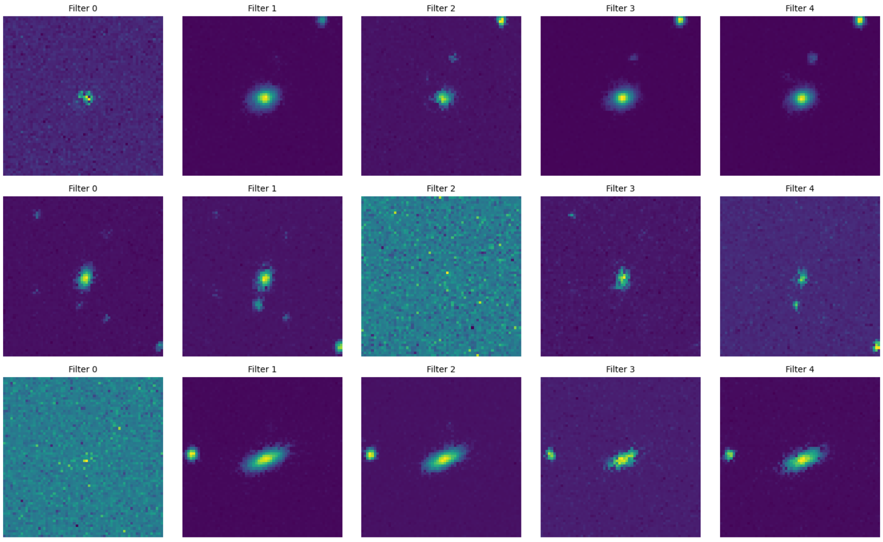

**Figure 2.5** Example of a cutout containing multiple detected sources, excluded from the final sample.

## 2.5 Detection and Removal of Multi-Object Cutouts

## 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* [6]. 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.*

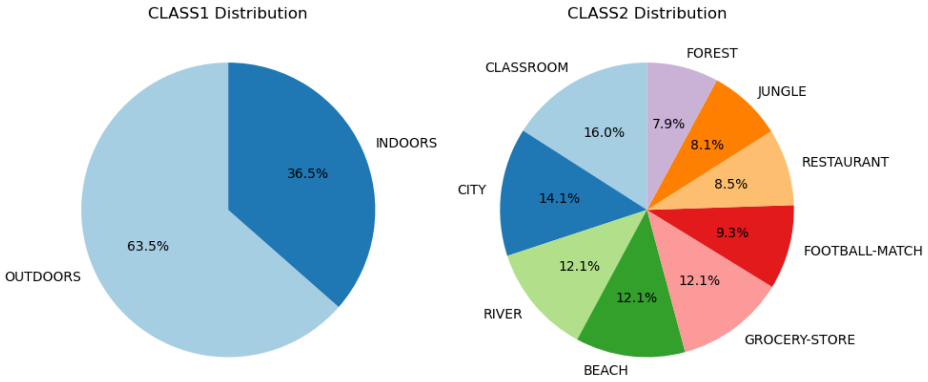

**Figure 3.1** CLASS1 (left) and CLASS2 (right) label distributions for the Scene dataset.

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.

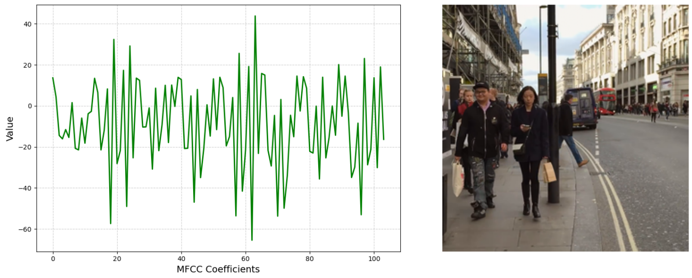

**Figure 3.2** MFCC plot of the audio and street-scene photo taken during the recording

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 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 [7].
- **Convolutional Neural Network (VGGNet12).** A 12-layer CNN architecture that excels at large-scale image feature extraction [8].
- **Gradient Boosting Machine (LightGBM).** An efficient implementation of gradient-boosted decision trees optimized for speed and memory [9].

## 3.3 Experimental Setup

### 3.3.1 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` [**KuhnJohnson2013**]. We then perform 5-fold cross-validation on the training set to estimate generalization error and tune hyperparameters [10, 11].

### 3.3.2 Preprocessing

- *Images:* Pixel values normalized to $[0, 1]$, then flattened for DT/LGBM or fed as 2D arrays to VGGNet12.
- *Spectra:* Any object with NaN flux values removed, yielding 11,179 gap-free spectra.
- *Early Fusion:* Concatenate image and spectral vectors into one feature vector.
- *Late Fusion:* Average photo-only and spec-only model predictions.

### 3.3.3 Hyperparameter Tuning

**DT:** grid search over `max_depth` $\in \{1, \ldots, 6\}$ with 5-fold CV, selecting the depth maximizing mean test $R^2$ [7]. **VGGNet12:** sweep over learning rate (`lr`) and fixed dropout=0.5, early stopping patience=30 [12, 13]. **LightGBM:** grid over `learning_rate` and `max_depth`, early stopping round=10 [**Probst2019**, 14].

## 3.4 Evaluation Metrics

We evaluate all models using:

- *Coefficient of Determination ($R^2$).* Variance explained [7].
- *Mean Absolute Error (MAE).* Average absolute deviation [7].
- *Root Mean Square Error (RMSE).* Quadratic penalty on large errors [7].
- *Normalized Median Absolute Deviation (NMAD).* $1.4826 \times \text{median}(|\epsilon - \text{median}(\epsilon)|)$ [15].

## 3.5 Multimodal Fusion Strategies

### 3.5.1 Early Fusion

Concatenate CNN feature vector (size $N_{\text{img}}$) with spectral vector (size $N_{\text{spec}}$) into one regressor input [16].

### 3.5.2 Late Fusion

Average independent predictions:

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

## 3.6 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 1:** Photo-only DT performance.

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

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

## 3.7 Convolutional Neural Network: VGGNet12

The VGGNet12 model stacks $3 \times 3$ convolutions, max-pooling, then three FC layers with dropout, fine-tuned from ImageNet [8].

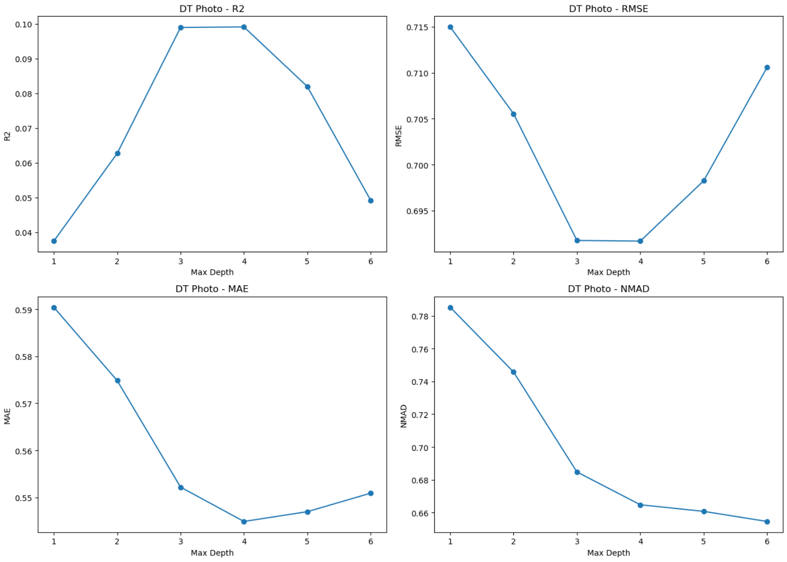

**Figure 3.3** DT on photographs: $R^2$, MAE, RMSE, and NMAD vs. max. tree depth. Best $d = 4$ (all except NMAD).

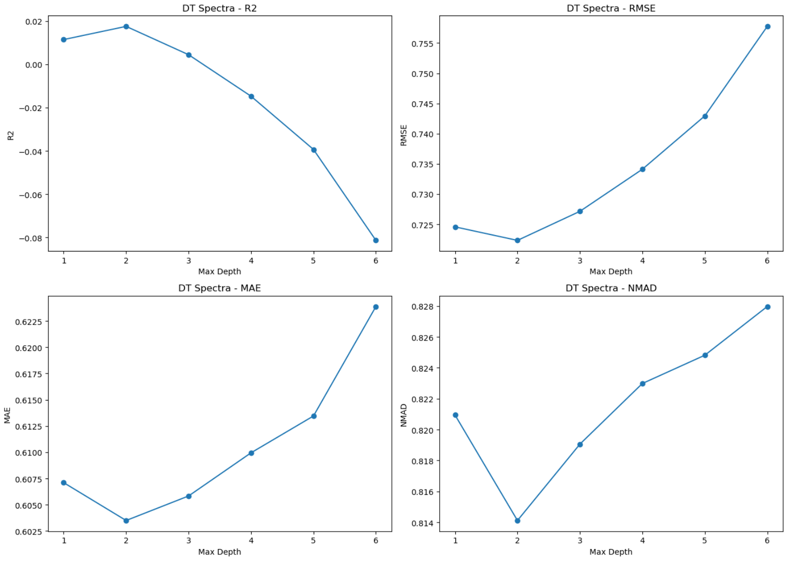

**Figure 3.4** DT on spectra: $R^2$, MAE, RMSE, and NMAD vs. max. tree depth. Best $d = 2$.

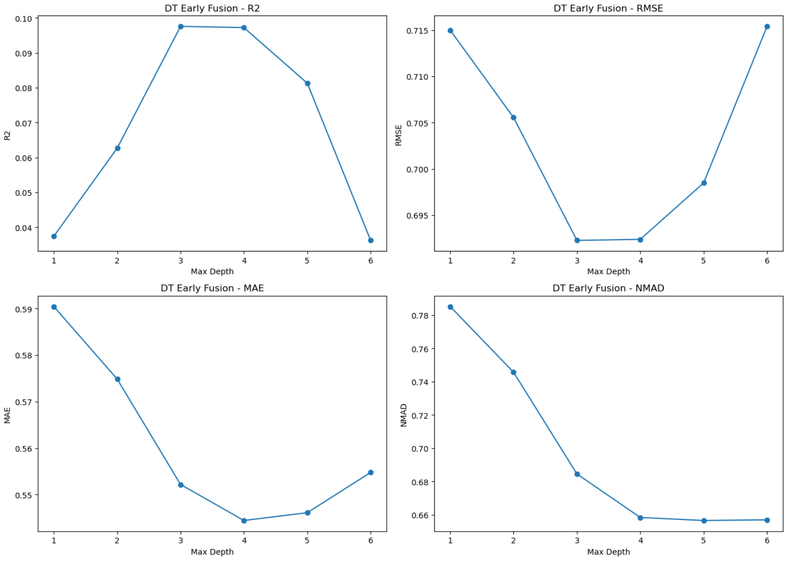

**Figure 3.5** DT early fusion: $R^2$, MAE, RMSE, and NMAD vs. tree depth. Best $d = 3$ by $R^2$.

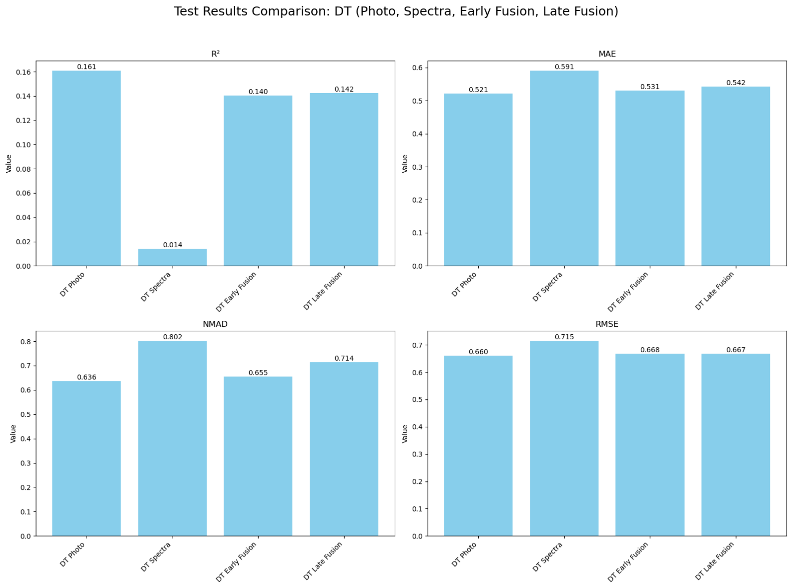

**Figure 3.6** DT: metric comparison across modalities (photo, spectra, early, late).

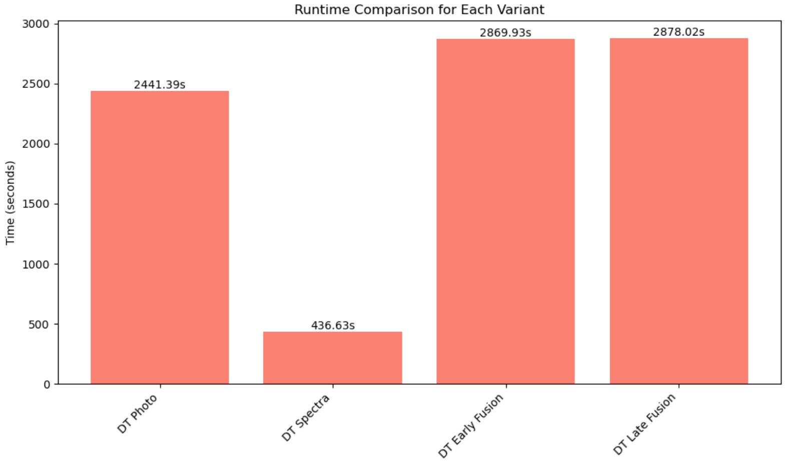

**Figure 3.7** DT: wall-clock runtime across modalities.

### 3.7.1 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 [17, 13, 12].

### 3.7.2 Training Curves: Photographs

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

### 3.7.3 Hyperparameter Sweep: Photographs

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

### 3.7.4 Training Curves: Spectra

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

### 3.7.5 Hyperparameter Sweep: Spectra

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

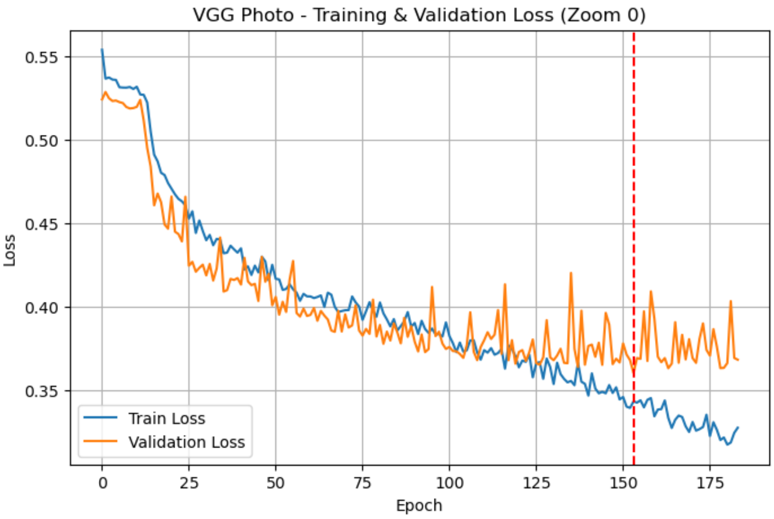

**Figure 3.8** VGGNet12 photo: training (blue) vs. validation (orange) loss per epoch; red dashed line marks lowest val. loss.

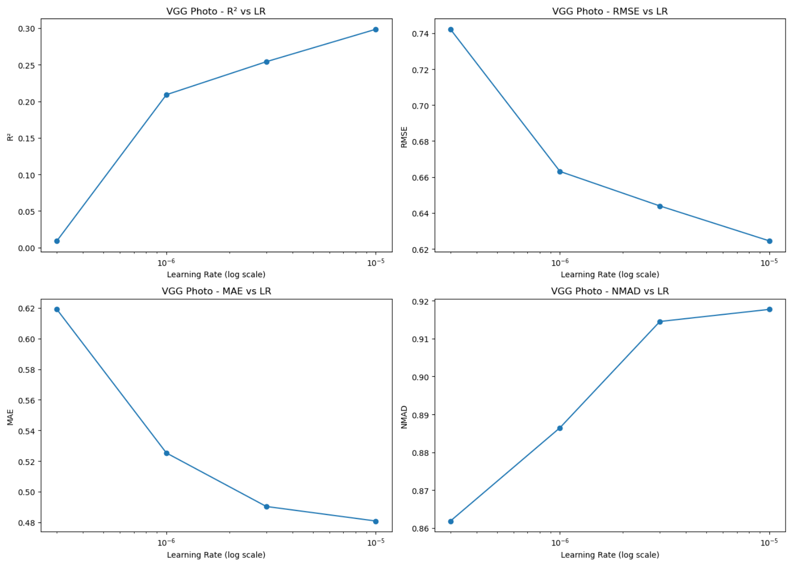

**Figure 3.9** VGGNet12 photo: $R^2$, MAE, RMSE, NMAD vs. learning rate.

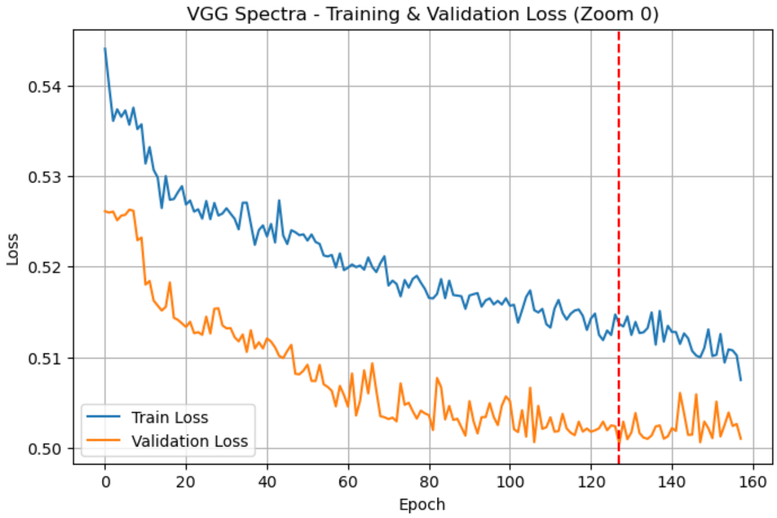

**Figure 3.10** VGGNet12 spectra: training vs. validation loss per epoch; red dashed line = best epoch.

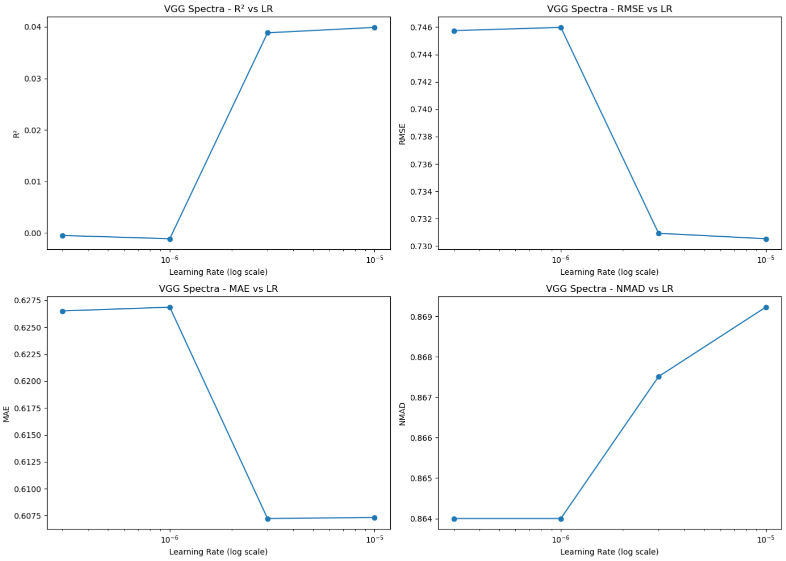

**Figure 3.11** VGGNet12 spectra: $R^2$, MAE, RMSE, NMAD vs. learning rate.

### 3.7.6 Training Curves: Early Fusion

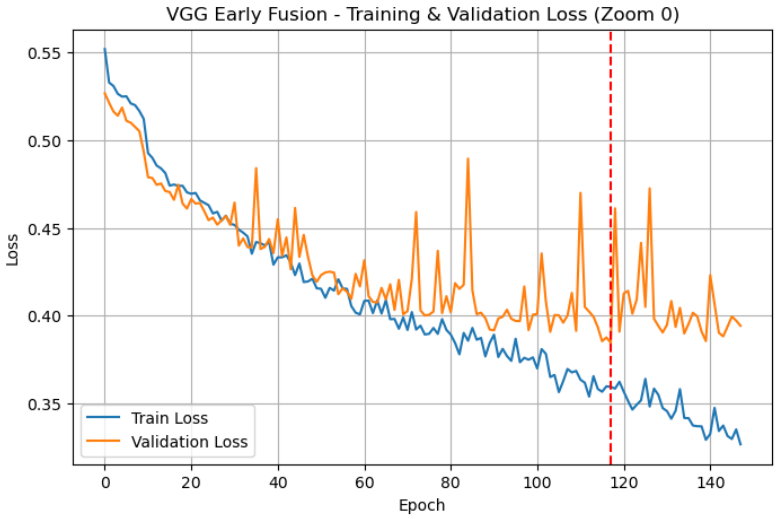

**Figure 3.12** VGGNet12 early fusion: training vs. validation loss; red dashed line = best epoch.

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

### 3.7.7 Hyperparameter Sweep: Early Fusion

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

### 3.7.8 Overall Metrics and Runtime

## 3.8 Gradient Boosting Machine: LightGBM

LightGBM grows trees leaf-wise with histogram-based splitting and optimizes RMSE with early stopping (10 rounds) [9, 14].

### 3.8.1 Architecture and Training Protocol

We minimize RMSE:

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

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

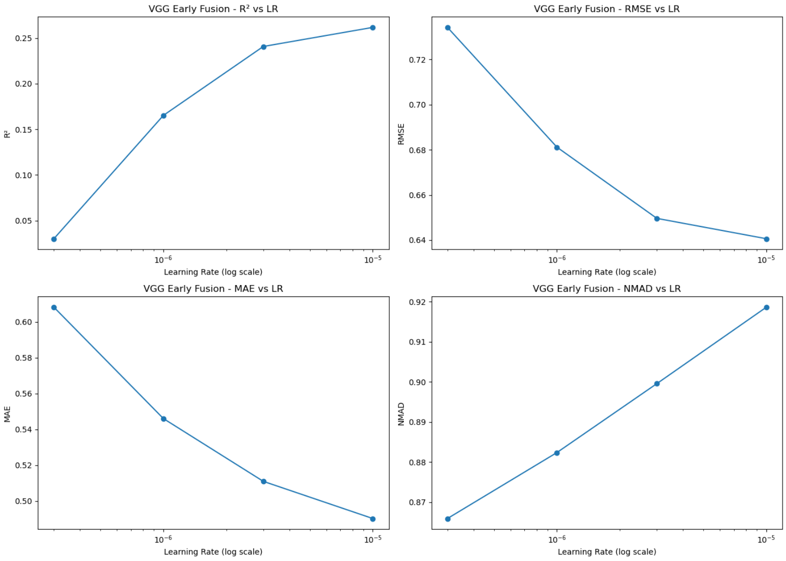

**Figure 3.13** VGGNet12 early fusion: $R^2$, MAE, RMSE, NMAD vs. learning rate.

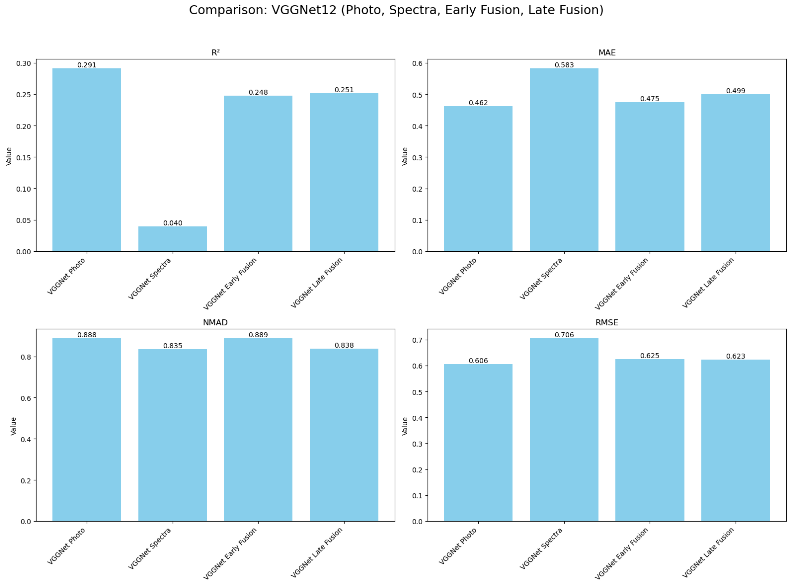

**Figure 3.14** VGGNet12: metric comparison across modalities.

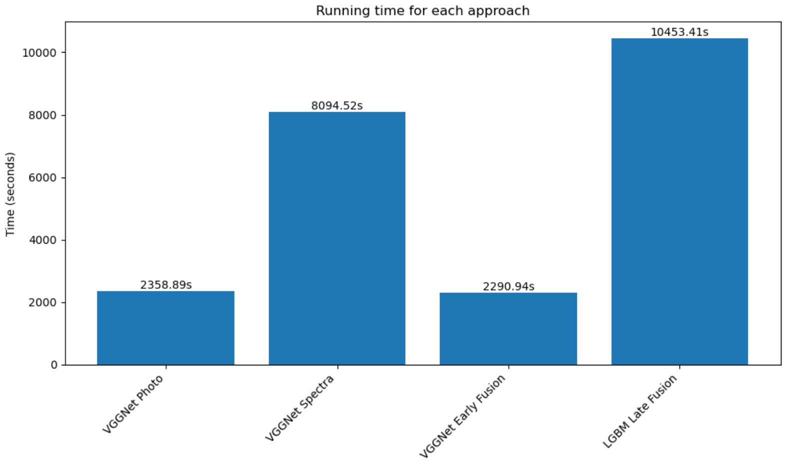

**Figure 3.15** VGGNet12: wall-clock runtime across modalities.

### 3.8.2 Training Curves: Photographs

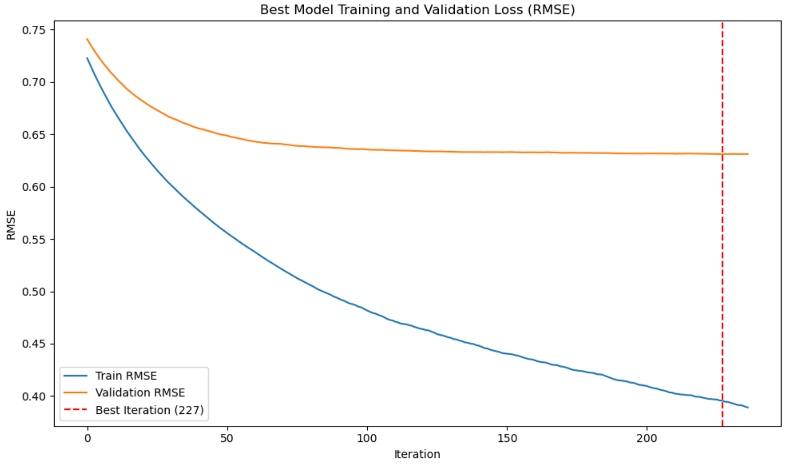

**Figure 3.16** LightGBM photo: training vs. validation RMSE per iteration; red dashed line = best iteration.

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

### 3.8.3 Hyperparameter Sweep: Photographs

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

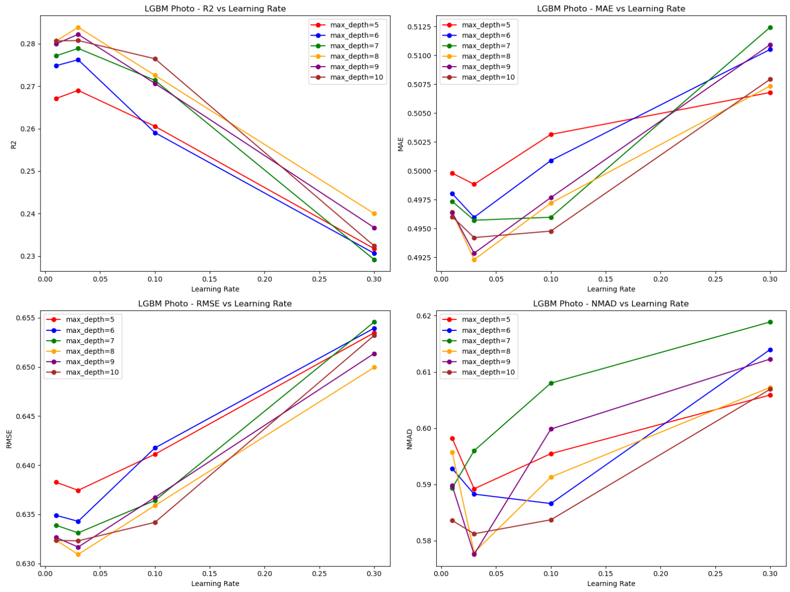

**Figure 3.17** LightGBM photo: $R^2$, MAE, RMSE, NMAD vs. learning rate & max_depth.

### 3.8.4 Training Curves: Spectra

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

### 3.8.5 Hyperparameter Sweep: Spectra

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

### 3.8.6 Training Curves: Early Fusion

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

### 3.8.7 Hyperparameter Sweep: Early Fusion

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

### 3.8.8 Overall Metrics and Runtime

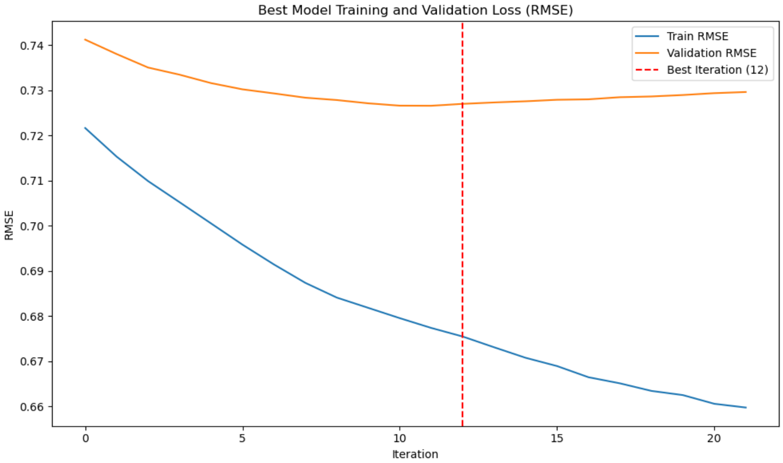

**Figure 3.18** LightGBM spectra: training vs. validation RMSE; red dashed line = best iteration.

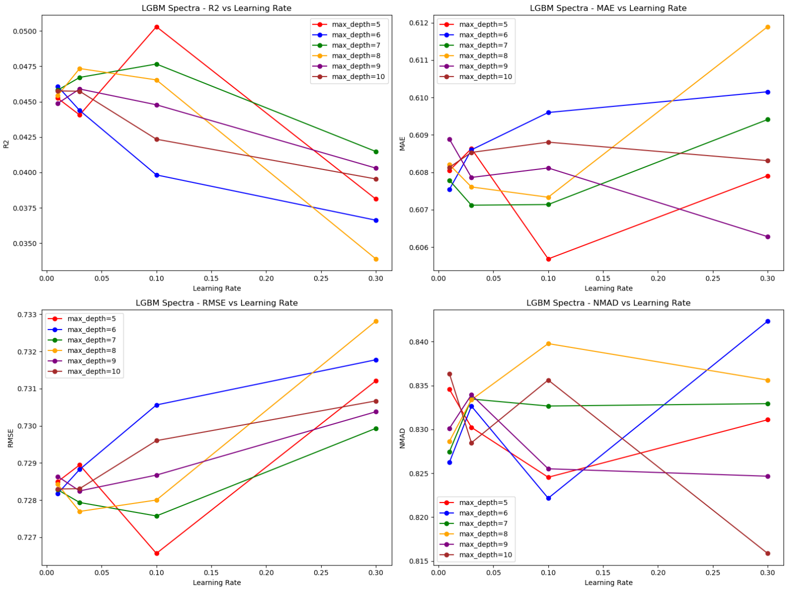

**Figure 3.19** LightGBM spectra: $R^2$, MAE, RMSE, NMAD vs. learning rate & max_depth.

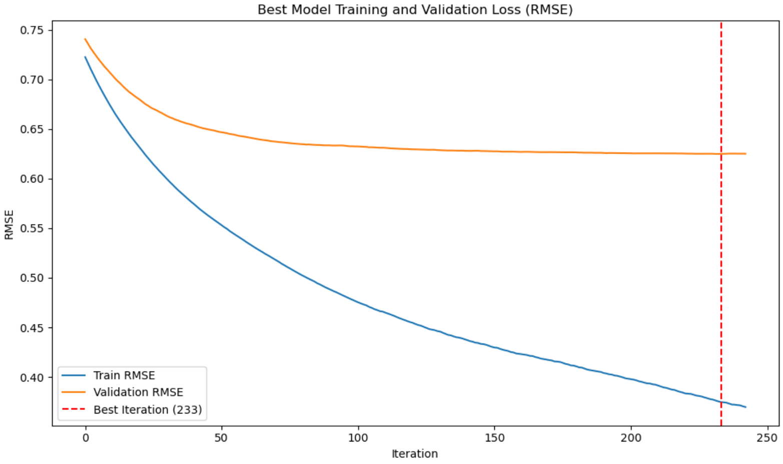

**Figure 3.20** LightGBM early fusion: training vs. validation RMSE; red dashed line = best iteration.

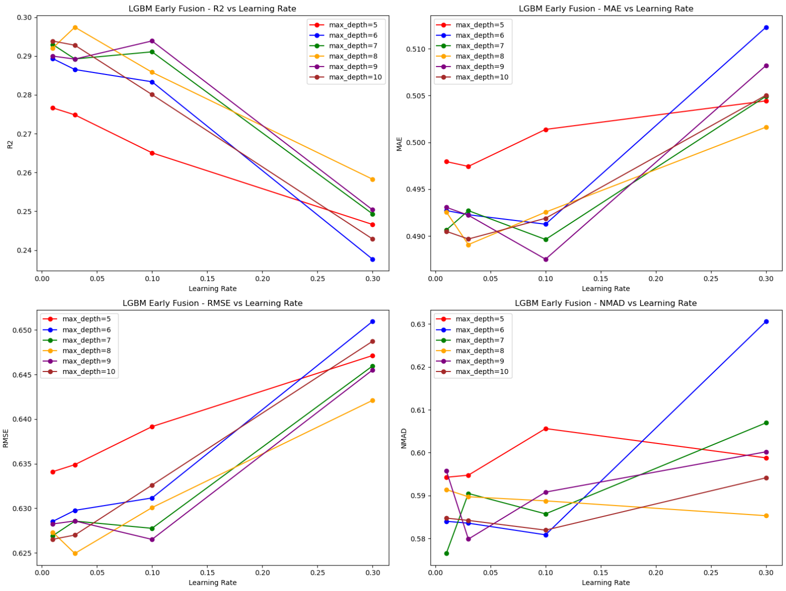

**Figure 3.21** LightGBM early fusion: $R^2$, MAE, RMSE, NMAD vs. learning rate & max_depth.

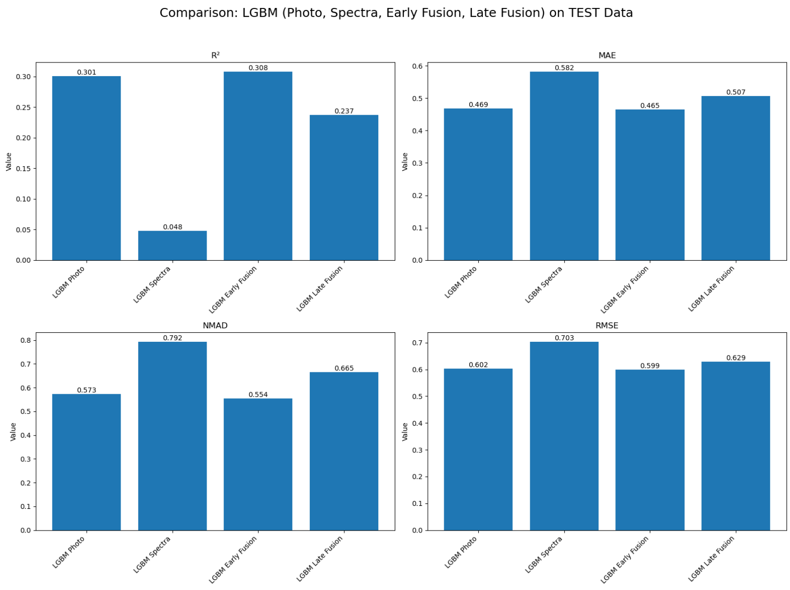

**Figure 3.22** LightGBM: metric comparison across modalities.

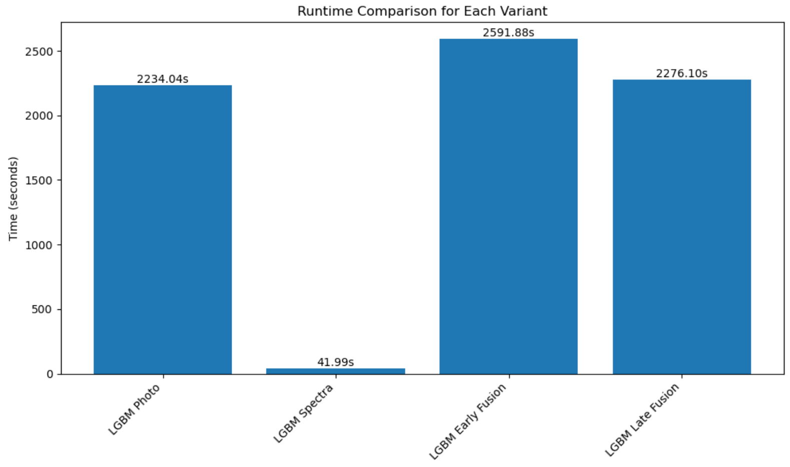

**Figure 3.23** LightGBM: wall-clock runtime across modalities.

## 3.9 Summary and Outlook

DT establishes a baseline, VGGNet12 adds visual feature depth, and LightGBM offers efficient boosting. Chapter **??** will synthesize all results and their astrophysical implications.

# Bibliography

1. 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.
2. 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.
3. MPA GARCHING. *Raw data*. 2007. Available also from: `https://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/raw_data.html`. [Online; accessed 2025-04-21].
4. MPA GARCHING. SDSS DR7 SFR documentation. *MPA Garching Web Resource*. 2007. Available also from: `https://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/sfrs.html`.
5. 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.
6. 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].
7. HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome H; FRIEDMAN, Jerome H. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.
8. SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 2014.

9. 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.

10. 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.

11. 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.

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

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

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

15. 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.

16. 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.

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