Machine Learning Experiments on Multi-Modal Astronomical Data

Farukh Rustamov
Supervised by: RNDr. Petr Škoda, CSc
CVUT
FIT

February 10, 2025

1 Introduction

1.1 Big Data in Astronomy

Modern astronomy deals with vast amounts of data collected from various telescopes and sky surveys. One of the largest and most influential surveys is the Sloan Digital Sky Survey (SDSS), which provides both multi-band images and spectroscopic data. The need for efficient processing and analysis of these datasets has led to the increasing use of machine learning techniques in astronomy.

The primary challenges associated with astronomical Big Data include:

- Data storage and accessibility: Managing petabytes of images and spectra.
- Automated data analysis: Traditional human-based classification is inefficient for large-scale datasets.
- Combining different modalities: Merging photometric and spectroscopic data for better scientific insights.

1.2 Machine Learning in Astronomy

Machine learning has been successfully applied in astronomy for various tasks, including:

- Classification of celestial objects: Distinguishing stars, galaxies, and quasars.
- Determination of redshift (z): Predicting galaxy distances based on photometry.
- Star formation rate (SFR) estimation: Inferring the rate of star formation using spectral features.
- Anomaly detection: Identifying rare astronomical events.

1.3 The Sloan Digital Sky Survey (SDSS)

The SDSS is one of the most comprehensive astronomical surveys, providing:

- Multi-band photometric images in five filters: u, g, r, i, and z.
- Spectra of millions of astronomical objects.
- Redshift and classification data for galaxies and quasars.

The five photometric filters (u, g, r, i, z) are designed to cover different parts of the electromagnetic spectrum:

- u: Near-ultraviolet (354 nm)
- g: Blue-green (477 nm)
- r: Red (623 nm)
- i: Near-infrared (762 nm)
- z: Infrared (913 nm)

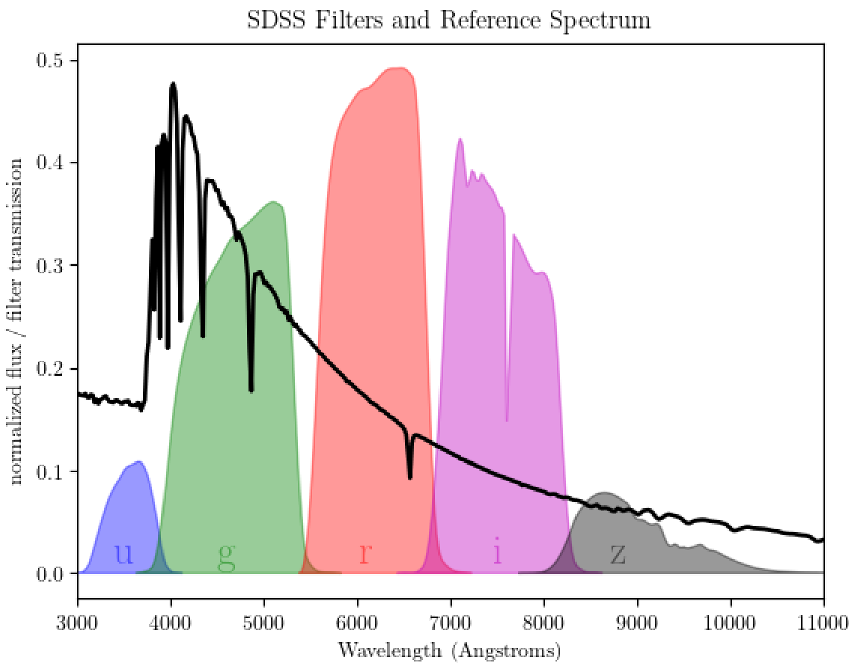


Figure 1: SDSS filter transmission curves

1.4 Understanding Spectroscopy

A spectrum is a representation of the intensity of light emitted by an object as a function of wavelength. Spectroscopic data provide crucial information about celestial bodies, including their chemical composition, temperature, motion, and distance.

Light from an astronomical object passes through a spectrograph, which disperses it into its component wavelengths. The resulting spectrum displays:

- Continuum emission, which reveals temperature and general energy distribution.
- Absorption and emission lines, which correspond to specific elements and molecules.
- Redshift (z), which helps determine an object's distance and velocity.

The redshift (z) is calculated using the formula:

$$z = \frac{\lambda_{observed} - \lambda_{rest}}{\lambda_{rest}} \tag{1}$$

where $\lambda_{observed}$ is the measured wavelength of a spectral line, and λ_{rest} is its known rest-frame wavelength.

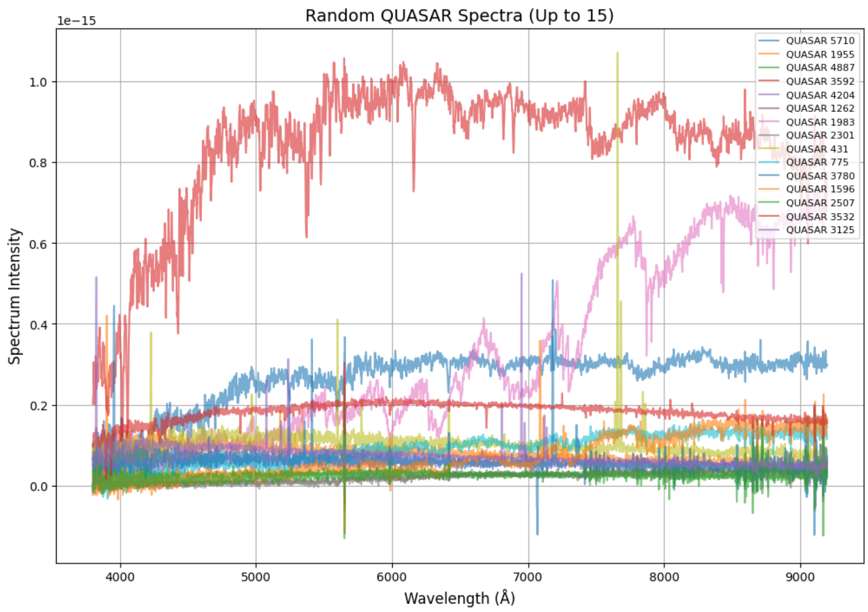
In SDSS, the combination of photometric filters and spectroscopy helps refine z estimates:

- Photometric redshifts use brightness variations in u, g, r, i, z bands.
- Spectroscopic redshifts rely on precise measurements of emission/absorption lines.

1.5 Spectra of Different Astronomical Objects

Different celestial objects have unique spectral signatures:

- **Stars:** Absorption lines from hydrogen (Balmer series), helium, and metals.
- **Galaxies:** Combination of stellar absorption lines and nebular emission lines (e.g., H-alpha, OIII).
- **Quasars:** Strong emission lines from highly ionized gas, broad line regions, and high redshift values.



1.6 Where to Find Spectra and Photometric Data

To access SDSS spectra and photometric images, the following sources are recommended:

- SDSS Main Site - Provides access to raw and processed data.
- SDSS SkyServer - Web-based interface to search and visualize spectra.
- SDSS Data Release 16 - Latest spectroscopic and photometric datasets.
- Legacy Survey - Complementary high-resolution images.

1.7 Goals of This Work

This research focuses on applying machine learning techniques to multi-modal astronomical data from SDSS. The key objectives are:

- Train ML models to predict redshift (z) using both images and spectra.
- Develop a model for star formation rate (SFR) estimation.
- Perform object classification (galaxy, star, quasar) using multimodal data.
- Compare the performance of multimodal ML models against unimodal approaches.

1.8 Structure of This Thesis

This thesis is structured as follows:

- Section 2 provides an overview of related works on ML applications in astronomy.
- Section 3 describes the dataset and preprocessing steps.
- Section 4 discusses the ML methods applied.
- Section 5 presents experimental results and comparisons.
- Section 6 concludes the research and suggests future work.