Introduction

Farukh Rustamov

March 15, 2025

1 General Description and Relevance of the Study

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). This study is motivated by the need to process and analyze large volumes of data and to develop optimal methods for the automated interpretation of observational data, a critical area in modern astronomy.

Originally, the entire SDSS dataset weighs hundreds of terabytes. Thanks to the HISS-CUBE system installed on the RCI cluster, a subset of the data has been set up, containing more than 150,000 objects. However, for proper statistical analysis and to minimize systematic errors, a filtering procedure is applied using the `FLAG` keyword. When `FLAG` is set to 0, the SFR estimation is considered reliable, and further calculations (e.g., SFR/M*) can be carried out without introducing subtle biases. As recommended on the dataset description page `https://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/raw_data.html`, we use only objects with `FLAG` = 0, reducing the dataset to about 11,000 objects (approximately 10,000 galaxies and 1,000 quasars).

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.

Special attention is given to the quality of the data. The images are available in four resolutions: 64x64, 32x32, 16x16, and 8x8 pixels. Similarly, the spectra are provided in four quality levels, allowing for an experimental selection of the optimal trade-off between data detail and processing speed. Further details regarding the quality levels can be added based on the original dataset specifications.

The processing and conversion of the raw data into a format suitable for machine learning were performed using HISS-CUBE. More details on this method are available in the publication at `https://www.sciencedirect.com/science/article/pii/S2213133721000172`.

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). Detailed information on the SFR parameter is available at `https://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/sfrs.html`. 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.

A comparative analysis of these approaches will help determine the method that yields the highest accuracy and whether the multimodal approach offers significant improvements over single-modality models.

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. This filtering minimizes potential distortions caused by alternative estimation methods.
2. **Quality of Images and Spectra.** The availability of multiple quality levels for both images and spectra allows for optimization of the data processing pipeline. However, determining the optimal resolution that balances processing speed and prediction accuracy is non-trivial.
3. **Multiple Objects in One Image.** There is a possibility that a single image may contain multiple objects. This issue must be addressed since overlapping signals can degrade the

performance of machine learning algorithms. Effective methods for automatic object detection and isolation need to be developed.

Addressing these challenges is crucial for improving the reliability and accuracy of the predictive models. The careful selection of objects and the optimization of data quality are key to minimizing systematic errors.

5 Comparative Analysis: The Scene Dataset Example

To preliminarily test the advantages of the multimodal approach, experiments were conducted on the Scene dataset, available at `https://www.kaggle.com/datasets/birdy654/scene-classificatio`
This dataset includes two types of data:

- **Images.** Frames extracted from videos that represent eight different types of environments.
- **Audio Data.** For each image, a set of MFCC (Mel-Frequency Cepstral Coefficients) attributes is provided, representing the audio characteristics corresponding to that frame.

Key aspects of the Scene dataset include:

- Images are extracted at regular intervals (e.g., one frame per second) and are accompanied by corresponding audio features.
- The dataset supports two classification tasks: binary (Indoors/Outdoors) and multi-class (8 types of environments).

Preliminary machine learning results on this dataset indicate that the multimodal approach significantly improves accuracy. For instance, decision tree experiments yielded:

- **Audio:** CLASS1 accuracy of 0.81 and CLASS2 accuracy of 0.66.
- **Combined:** CLASS1 accuracy of 0.97 and CLASS2 accuracy of 0.92.

Additionally, neural network experiments showed:

- **Audio:** CLASS2 test accuracy of 0.94.
- **Images:** CLASS2 test accuracy of 0.99.
- **Combined:** CLASS2 test accuracy of 0.99.

These results support the hypothesis that integrating modalities can enhance prediction accuracy, encouraging the application of similar techniques to the SDSS data.

6 Objectives and Tasks

The primary objective of this thesis is to develop an optimal methodology for predicting the star formation rate (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 to address the issue of multiple objects per image.
3. Investigate the impact of different quality levels of images and spectra on prediction accuracy and determine the optimal resolutions for each modality.
4. Compare the effectiveness of models using single modalities (either images or spectra) with multimodal approaches that combine both data types.
5. Conduct a comparative study using the Scene dataset as a preliminary experiment and then adapt the findings to the analysis of SDSS data.

Achieving these tasks will not only improve the accuracy of SFR predictions but also optimize computational efficiency, which is crucial when dealing with large-scale astronomical datasets.

7 Terminology and Illustrations

Understanding key terminology and methods is essential for interpreting the data and the results obtained from the machine learning experiments. This section provides an overview of fundamental concepts and illustrates them with relevant figures.

7.1 Spectra and Spectral Analysis

7.1.1 Definition of a Spectrum

A spectrum in astronomy represents the dependence of an object's emitted intensity on wavelength (or frequency). When light passes through a prism or a diffraction grating, it is dispersed into its constituent wavelengths, ranging from the ultraviolet to the infrared. Specialized spectrographs attached to telescopes are used to record these spectra.

7.1.2 Why Spectral Analysis Is Needed

Spectral analysis is a powerful tool in astronomy because:

- **Chemical Composition:** Each element, such as hydrogen, helium, or iron, exhibits characteristic spectral lines. By analyzing these lines, astronomers can determine the chemical makeup of celestial objects and infer physical properties like temperature and density.

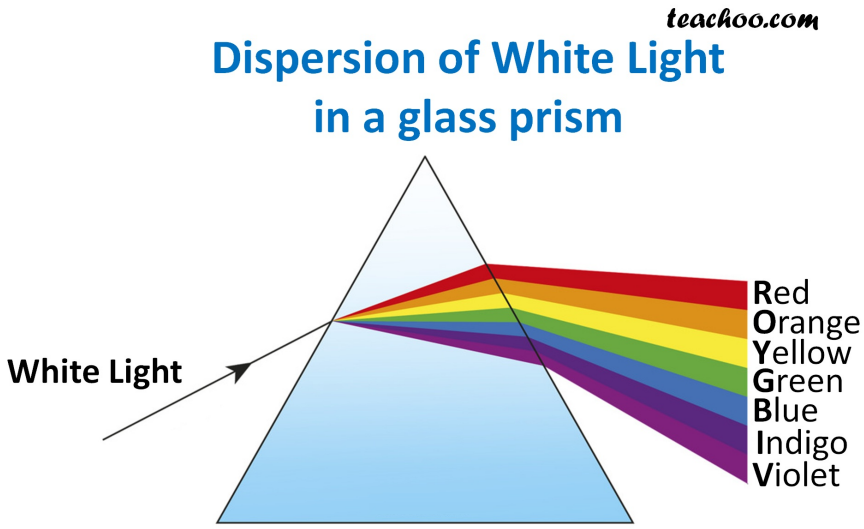

Figure 1: White-light decomposition through a prism.

- **Velocity Measurements:** Shifts in the wavelengths of spectral lines provide insights into the motion of celestial objects.
- **Physical Conditions:** The presence and intensity of emission or absorption lines can indicate conditions such as ionization levels and star formation activity.

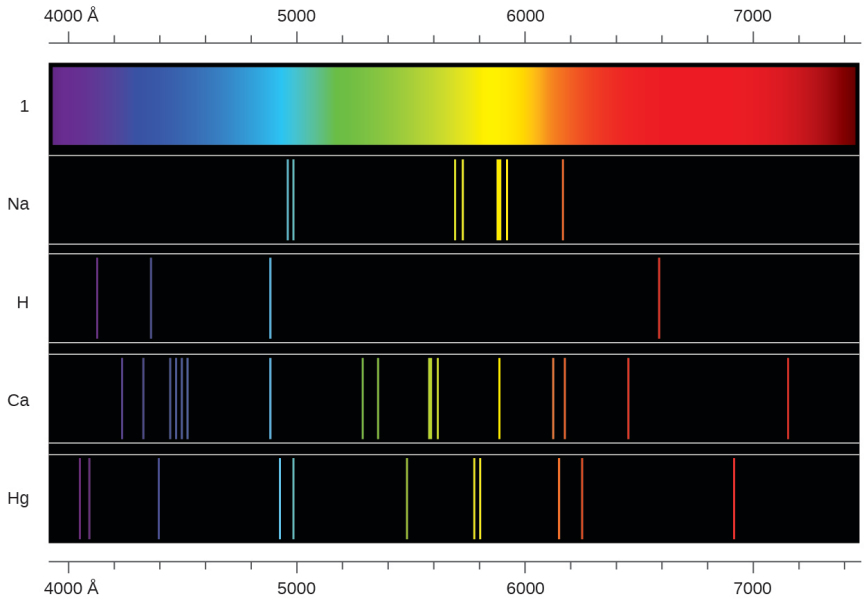

Figure 2: Example of atomic spectral lines for different elements.

7.1.3 Typical Types of Spectra

- **Stellar Spectrum:** Characterized by prominent absorption lines, which are indicative of the elements present in a star's atmosphere.
- **Nebula or Gas Cloud Spectrum:** Dominated by bright emission lines produced by ionized gas.

- **Galaxy Spectrum:** A composite spectrum that includes the integrated light of billions of stars as well as emission from interstellar gas.

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

One of the most influential surveys in modern astronomy is the Sloan Digital Sky Survey (SDSS). It provides both photometric and spectroscopic data for millions of celestial objects. The SDSS photometric system is built around five broadband filters:

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

These filters allow astronomers to construct a broadband Spectral Energy Distribution (SED) for each object, which is critical for understanding its physical properties.

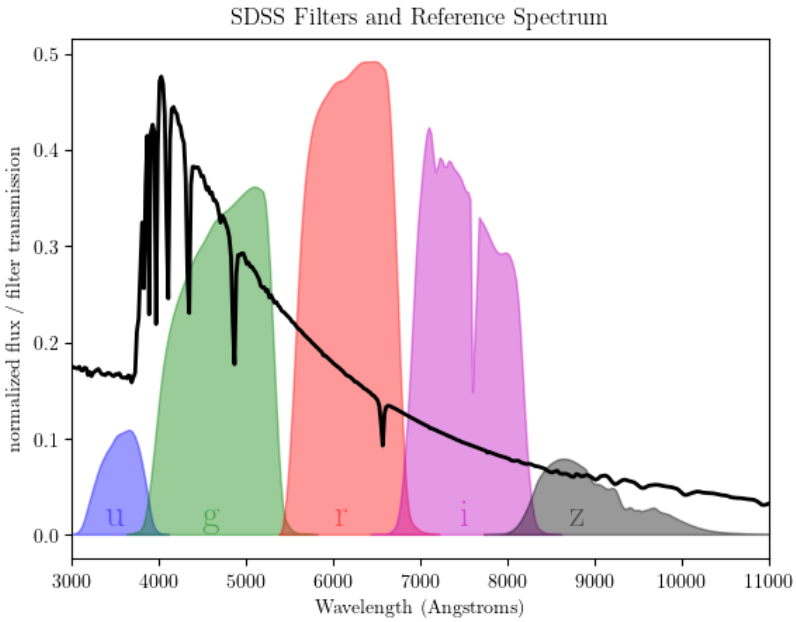

Figure 3: Transmission curves of the SDSS *u, g, r, i, z* filters.

7.3 Star Formation Rate (SFR)

7.3.1 What Is SFR

The Star Formation Rate (SFR) quantifies the rate at which a galaxy forms new stars, typically expressed in solar masses per year ($M_\odot\,\mathrm{yr}^{-1}$).

7.3.2 How SFR Is Determined

A common method to estimate SFR is by measuring the intensity of specific emission lines (e.g., Hα) in a galaxy's spectrum. Since young, massive stars ionize the surrounding hydrogen gas, the Hα luminosity is directly related to the current star formation activity:

$$\mathrm{SFR} \propto L(\mathrm{H}\alpha), \tag{1}$$

with a standard calibration given by:

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

Other methods may use ultraviolet or infrared measurements, depending on data availability and dust obscuration.

7.3.3 Why SFR Matters

- **Galactic Evolution:** SFR is a key parameter in understanding how galaxies evolve over time.
- **Energy Output:** High SFRs indicate intense star formation, which can have significant effects on the interstellar medium and the future evolution of a galaxy.

8 Conclusion of the Introduction

In summary, this thesis explores the prediction of the star formation rate (SFR) in astronomical objects using SDSS data. The study leverages both photometric images and spectroscopic data. By selecting objects with `FLAG` $= 0$, we ensure the reliability of the data, while the availability of multiple quality levels allows for an optimal balance between processing speed and accuracy.

The preliminary analysis on the Scene dataset has demonstrated that a multimodal approach significantly enhances prediction accuracy, motivating the application of similar methods to the SDSS data. Furthermore, understanding the underlying terminology, such as the nature of spectra, the function of the SDSS *u, g, r, i, z* filters, and the methods for determining SFR, is essential for interpreting the data and the results of the machine learning experiments.

Future sections of this work will detail the methodology, experimental results, and comparative analysis of various machine learning algorithms in the context of predicting the star formation rate.

9 Outline of Future Work and Machine Learning Approach

In the next phase of this research, we plan to develop a comprehensive machine learning framework for SFR prediction that leverages both imaging and spectroscopic data. Below is an outline of the intended research structure and experimental pipeline.

9.1 Data Acquisition and Preparation

- **Data Exploration:** Begin by thoroughly examining the SDSS dataset to understand the characteristics of the photometric images and spectroscopic data. Visualize samples from each modality and assess data quality.
- **Addressing Multiple Objects:** Investigate the issue of multiple objects per image. Possible approaches include removing extraneous objects or cropping images to retain only the central region.
- **Statistical Analysis of SFR:** Analyze the distribution of the SFR (represented by the `AVR` column) to understand its behavior and range, which is crucial for subsequent regression tasks.

9.2 HiSS-Cube Implementation and Data Structuring

- **HiSS-Cube Overview:** Utilize the Hierarchical Scientific Storage Cube (HiSS-Cube) built on HDF5 for efficient data management. This structure enables seamless association between five-band images, full spectroscopic data, and computed SFR values.
- **Data Organization:** Organize the dataset in a hierarchical format, ensuring pixel-to-pixel and wavelength-to-wavelength correspondence, along with metadata such as mass and SFR.

9.3 Baseline Models: Single-Modality Experiments

- **Spectroscopic Baseline:** Develop a model using a decision tree and then a hybrid network (combining convolutional and recurrent layers) to extract spectral features. Analyze performance using regression metrics.
- **Imaging Baseline:** Implement a convolutional neural network (CNN) tailored to the five-band (u, g, r, i, z) images. Experiment with both individual channel processing and combined-band approaches.

9.4 Fusion Strategies and Neural Network Architectures

9.4.1 Early Fusion (EarlyFusionNet)

- **Description:** Directly concatenates PCA-processed image (100 components) and spectral features (50 components) into a unified input vector. This combined input is processed through a fully connected neural network consisting of two hidden layers (128 and 64 neurons), ReLU activation functions, and dropout regularization (rates of 0.3 and 0.2).
- **Advantage:** Simplicity and minimal computational overhead.
- **Limitation:** Potential loss of modality-specific feature extraction since modalities are merged at an early stage, possibly causing information dilution.
- **Reference:** Early versus Late Fusion

9.4.2 Intermediate Fusion (FusionNet)

- **Description:** Utilizes separate branches for imaging and spectral modalities:
  - Imaging data is processed via a Convolutional Neural Network (CNN) optimized to extract spatial features.
  - Spectral data passes through a two-layer fully connected neural network.

  Extracted features are concatenated at an intermediate layer and subsequently processed by a fusion layer consisting of fully connected layers with dropout.
- **Advantages:**
  - Captures modality-specific information separately before integrating them.
  - Effectively balances complexity and performance.
  - Enhanced by regularization methods (dropout, early stopping) to prevent overfitting.
- **Limitation:** Slightly more computational complexity compared to Early Fusion.
- **Reference:** Multimodal Machine Learning Survey

9.4.3 Late Fusion (BranchNet)

- **Description:** Employs two completely independent neural networks for images and spectra, each trained separately to predict SFR. Final predictions from both models are averaged to provide the final regression output.
- **Advantage:** Allows each model to specialize fully in capturing unique modality characteristics without interference.
- **Limitation:** May miss subtle interactions between modalities due to its independence during feature extraction.

- **Reference:** Early versus Late Fusion

9.5 Regularization and Prevention of Overfitting

Significant enhancements were applied to mitigate the issue of overfitting prevalent in earlier implementations. Specifically, these improvements include:

- **Dropout Regularization:** Randomly disables neurons during training, ensuring that the neural networks do not overly rely on specific pathways or features. Dropout rates (0.2–0.5) substantially enhance generalization performance. [Dropout Reference]
- **Early Stopping:** Implemented to terminate the training process automatically when validation loss ceases to decrease. This method has proven effective to prevent overfitting by ensuring the network retains its best generalization capabilities. [Early Stopping Reference]
- **Proper Data Preprocessing:** PCA and standard scaling parameters were computed exclusively on the training dataset, avoiding information leakage. This procedure ensures unbiased evaluation and robust generalization to unseen data.
- **Dimensionality Reduction (PCA):** PCA reduces noise, computational complexity, and input dimensionality (100 PCA components for imaging and 50 for spectra), facilitating better learning outcomes.

9.6 Experimental Results and Analysis

Experimental validation conducted with controlled setups yielded the following comparative metrics:

- **Early Fusion:** MSE ≈ 0.382, $R^2 \approx 0.23$
- **Intermediate Fusion:** MSE ≈ 0.368, $R^2 \approx 0.26$ (best performing)
- **Late Fusion:** MSE ≈ 0.382, $R^2 \approx 0.23$

Contrary to previous experiments indicating the superiority of Late Fusion, the refined Intermediate Fusion architecture achieved the best results in the current experiments. This improvement is primarily attributed to:

- Application of CNN architecture to imaging data, significantly enhancing spatial feature extraction.
- Robust regularization via dropout and early stopping, substantially improving generalization capability.

9.7 Comparative Advantages of Proposed Approach

The improved neural network approaches presented in this research surpass previous implementations due to the following critical advancements:

- **Specialized Modality Processing:**
  - CNN layers specifically optimized for imaging data are superior in capturing complex spatial patterns compared to purely fully connected layers.
  - Dedicated MLP layers for spectral data enhance extraction of modality-specific features.
- **Better Generalization:**
  - Early stopping, dropout regularization, and PCA preprocessing collectively strengthen the models against overfitting.
  - The improved data preprocessing strategy (training-exclusive PCA fitting) avoids data leakage, enhancing model reliability.
- **Flexibility and Scalability:** Modular network design enables efficient extension and scalability, suitable for larger datasets anticipated from future astronomical surveys.

9.8 Conclusion

These experiments demonstrate clearly that a carefully designed Intermediate Fusion approach, incorporating CNN layers for image modality and explicit regularization strategies, outperforms simpler fusion techniques for multimodal regression tasks in astronomy. Future work should explore more sophisticated architectures (e.g., ResNet and attention mechanisms), as well as contrastive learning approaches, to further enhance predictive performance and robustness.

9.9 Multi-Modal Fusion Approaches

- **Fusion Strategies:** Investigate various fusion techniques, including:
  - **Early Fusion:** Integrating data at the input level.
  - **Intermediate Fusion:** Merging feature representations extracted separately from each modality.
  - **Late Fusion:** Combining decisions or predictions from independent models.
- **Hybrid Architecture:** Develop an architecture that processes the five SDSS bands using a modified ResNet for imaging, alongside a specialized network for spectral features (designed to capture emission line strengths and continuum patterns).
- **Contrastive Learning:** Apply contrastive learning techniques to generate a joint embedding space from both modalities, with the hypothesis that this joint space captures complementary information that improves SFR regression.

9.10 Experimental Design and Validation

- **Controlled Experiments:** Establish a testing framework with carefully controlled experiments comparing single-modality models with the proposed multi-modal fusion approaches.
- **Evaluation Metrics:** Use standard regression metrics such as RMSE, MAE, and R^2 to evaluate performance. In addition, validate predictions against theoretical models of galaxy evolution.
- **Cross-Attention Mechanisms:** Explore fusion approaches that incorporate cross-attention to allow each modality to highlight relevant features in the other.

9.11 Comparative Analysis and Future Directions

- **Model Comparison:** Compare the performance of conventional machine learning models (e.g., decision trees) with deep learning fusion architectures.
- **Scalability and Efficiency:** Evaluate the computational efficiency of each approach, particularly considering the large-scale nature of astronomical data.
- **Extension to Other Parameters:** Outline plans to extend the multimodal approach to predict additional astronomical parameters beyond SFR in future work.

This research framework builds upon recent advances in multi-modal datasets and contrastive learning applied to astronomical data. The methodologies and experiments proposed here aim to demonstrate that integrating imaging and spectroscopic data through advanced fusion techniques significantly improves the prediction of star formation rates. Furthermore, the HiSS-Cube data structure will serve as a robust backbone for efficient data management and processing, paving the way for future astroinformatics applications.

Future work will also investigate the potential for applying these techniques to upcoming large-scale surveys, which will provide unprecedented volumes of multimodal astronomical data.