Galaxy Morphology Classification Using Deep Learning

A dissertation submitted in partial fulfilment for the degree of

Master of Technology

in

Computer Science

by

Dipanwita Kundu Roy

Roll No.: **CS2414**

under the supervision of

Prof. Sarbani Palit

Computer Vision and Pattern Recognition Unit (CVPRU)

Indian Statistical Institute, Kolkata

June, 2026

CERTIFICATE

This is to certify that the dissertation titled **Galaxy Morphology Classification Using Deep Learning** submitted by Dipanwita Kundu Roy to the Indian Statistical Institute, Kolkata, fulfills the requirements for the degree of Master of Technology in Computer Science. The work presented in this dissertation is an authentic and original contribution conducted under my supervision and guidance. I confirm that this dissertation adheres to all the necessary academic standards and regulations of the institute.

Dr. Sarbani Palit
CVPR Unit
Indian Statistical Institute
Kolkata - 700108
India

Acknowledgement

I would like to express my gratitude to Dr. Sarbani Palit, my supervisor at the Computer Vision and Pattern Recognition Unit of the Indian Statistical Institute in Kolkata. Her expert guidance, unwavering support, and inspiring insights have been invaluable throughout my academic journey. Dr. Palit's extensive knowledge and innovative ideas have significantly contributed to my understanding of various topics and have been crucial in shaping my research skills and research aptitude.

I am very thankful to Ankita Sarkar, Senior Research Fellow at the Indian Statistical Institute, for his essential assistance in acquiring the datasets, notes and necessary materials for this study. Her continuous input of ideas and continuous support have played an important role in the successful completion of this project. After various discussion related to topic we came with some improvement in our work.

I also extend my gratitude to all the faculty members at the Indian Statistical Institute for their invaluable advice, insights, and teachings, which have provided a vital perspective to my research work.

I am thankful to all my friends for their ongoing assistance and motivation. My appreciation goes out to everyone who has contributed to my personal and academic growth, even if I have not explicitly mentioned them here.

Declaration

I, **Dipanwita Kundu Roy**, with Roll No. **CS2414**, hereby declare that the material presented in the dissertation titled **Galaxy Morphology Classification Using Deep Learning** represents original work carried out by me for the degree of **Master of Technology in Computer Science** at the **Indian Statistical Institute, Kolkata**.

Furthermore, I affirm that no sections of this report have been sourced or copied from external references without proper attribution. I am aware that any instances of plagiarism or the use of unacknowledged materials from third parties will be treated with the utmost seriousness and consequences.

Dipanwita Kundu Roy
M.Tech (CS), CS2414
Indian Statistical Institute

Abstract

Galaxy morphology is the study of the shape and visual appearance of galaxies, such as spiral, smooth, edge-on, and other morphological types. Morphological classification plays an important role in understanding how galaxies form and evolve over cosmic time. Most existing machine learning approaches for galaxy morphology classification rely solely on RGB galaxy images, which primarily capture spatial information and lack the physical spectral context of galaxies. In contrast, astronomical spectral datacubes contain rich information across multiple wavelengths, providing insights into the internal and physical properties of galaxies. However, such spectral observations are available for only a limited number of objects. Motivated by the availability of spectral information during training, this work investigates whether spectral knowledge can be used to improve morphology classification when only RGB images are available during testing. Different embedding techniques, including Siamese Networks, Supervised Autoencoders, and channel-based architectures, are explored to learn meaningful latent representations from spectral datacubes. These embeddings are then integrated with corresponding RGB galaxy images using multimodal deep learning frameworks for effective feature learning and classification. Experimental results demonstrate that incorporating spectral embeddings during training can guide the learning process and improve galaxy morphology classification performance using image-only inputs at inference time.

Keywords: Galaxy Morphology Classification, Datacube Embedding, Supervised Autoencoder, Siamese Network, Multimodal Learning, Gated Network

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Literature Survey

Galaxy morphology classification has remained one of the most important research areas in astronomy because galaxy structure and morphology look into galaxy formation, evolution, stellar population, and environmental interactions. Early classification systems were based on visual inspection, among which the Hubble classification scheme remains one of the most influential frameworks. Galaxies were mainly categorized into elliptical, spiral, barred spiral, and irregular types based on their visual appearance.

With the development of large astronomical surveys such as the Sloan Digital Sky Survey (SDSS), Legacy Survey of Space and Time (LSST), James Webb Space Telescope (JWST), and Galaxy Zoo, massive amounts of galaxy image data became available, making manual classification increasingly difficult. The Galaxy Zoo project introduced large-scale citizen-science-based morphological labeling and provided detailed morphology annotations for more than 300,000 galaxies from SDSS [2]. These datasets became standard benchmarks for automated galaxy morphology classification research.

Traditional automated methods relied on features such as concentration index, surface brightness profile, Gini coefficient, and moment-based features for classification. However, these methods were limited by feature engineering and those also struggled to generalize across diverse galaxy morphologies.

Recent advances in deep learning significantly improved galaxy morphology classification performance. Convolutional Neural Networks (CNNs), Residual Networks, DenseNet, EfficientNet, Capsule Networks [4], and attention-based architectures have been widely applied for extracting hierarchical spatial features from galaxy images. Although image-based methods achieved high accuracy, RGB galaxy images alone often fail to capture complete physical and spectral properties of galaxies. Astronomical datacubes [1] obtained from Integral Field Spectroscopy (IFS) preserve both spatial and spectral information and provide deeper insights into stellar population, metallicity, star formation activity, and galaxy evolution. However, the extremely high dimensionality of datacubes makes direct utilization computationally challenging.

To address this issue, recent studies explored representation learning and embedding techniques for astronomical datacubes. Contrastive learning [1] have been used to generate meaningful latent embeddings capable of separating galaxies according to morphology, stellar mass, metallicity, and star formation properties. These studies demonstrated that learned embedding spaces can effectively capture physical galaxy characteristics from spectral data.

1.2 Our Proposed Work

The proposed work focuses on how to improve galaxy morphology classification by integrating spectral information from datacubes with corresponding RGB galaxy images. Different embedding generation techniques, including Siamese Networks, Supervised Autoencoders, and channel-based architectures, are explored to learn compact latent representations from datacubes.

The generated embeddings are combined with corresponding RGB galaxy cutouts using different multimodal deep learning approaches. Various fusion strategies, gated networks, and attention mechanisms are tried to effectively utilize both spatial and spectral information for galaxy classification. Comparative analysis is performed to evaluate the effectiveness of different embedding and multimodal architectures in improving classification performance.

Chapter 2

Dataset Description

2.1 eCALIFA Datacube Dataset

The primary spectral dataset used in this work is the extended data release (eDR) of the CALIFA survey, commonly referred to as the eCALIFA dataset [1]. The observations were acquired using the 3.5-m telescope at Calar Alto Observatory [15] with the PMAS/PPak integral field spectrograph and the V500 grating setup. The spectra cover a wavelength range of 3745–7500 Å with a spectral resolution of approximately $R \sim 850$. The instrument provides a hexagonal field of view of 74×64 arcsec^2 using 331 science fibers, each having a diameter of 2.7 arcsec.

To improve the quality of the observations, dithering techniques and image reconstruction algorithms were applied during preprocessing. The data were reconstructed using flux-conservative interpolation and resampled into regularly sampled datacubes. These preprocessing techniques significantly improved the spatial resolution and image quality of the dataset.

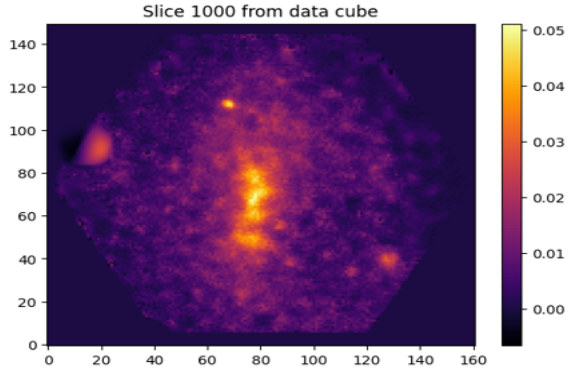
The eCALIFA dataset contains nearby galaxies within a redshift range of 0.0005 to 0.08, with most galaxies located below $z < 0.03$. After visual inspection and quality filtering, only high-quality galaxy datacubes were selected for the present study. These datacubes preserve both spatial and spectral information across multiple wavelengths, so that those become useful for studying galaxy morphology, stellar population, metallicity, and galaxy evolution.

The eCALIFA dataset contains galaxies belonging to a variety of morphological classes. The morphology labels are provided by the CALIFA survey based on visual inspection and include **BCD, E0–E7, I, S0, S0a, Sa, Sab, Sb, Sbc, Sc, Scd, Sd, and Sdm**. Although these classes are similar to those in the Hubble sequence, eCALIFA uses its own morphology catalog rather than directly adopting the standard Hubble classification system. The dataset therefore covers a wide range of galaxy types, from elliptical and lenticular galaxies to spiral and irregular systems.

2.2 RGB Galaxy Images from SDSS DR10

For each galaxy present in the eCALIFA dataset, the corresponding RGB galaxy cutout image was collected from the Sloan Digital Sky Survey (SDSS) Data Release 10 (DR10). DR10 provides calibrated optical imaging and spectroscopic observations collected through the SDSS and BOSS survey projects.

A total of 786 RGB galaxy cutouts were successfully matched with their corresponding eCALIFA datacubes. These corresponding RGB preserve the same morphology classes as their paired eCALIFA datacubes.



(a) Spectral slice pseudo-color visualization of datacube

(b) RGB image of corresponding datacube

Figure 2.1: Visualization of the datacube and its corresponding RGB galaxy image.

2.3 Galaxy Zoo 2 Dataset

Galaxy Zoo 2 (GZ2) [2] is an extension of the original Galaxy Zoo project designed to provide detailed morphological classification of galaxies collected from the Sloan Digital Sky Survey (SDSS). The dataset was generated through a large-scale citizen science initiative in which volunteers classified galaxies using a hierarchical decision-tree-based questionnaire.

Unlike the original Galaxy Zoo dataset, GZ2 provides fine-grained morphological annotations including spiral arms, galactic bars, bulges, edge-on structures, and ellipticity information. The complete dataset contains morphological classifications for 304,122 galaxies collected from SDSS observations.

The Galaxy Zoo 2 dataset serves as one of the most important benchmark datasets for automated galaxy morphology classification research.

As part of the GZ2 data release, a shorthand morphology representation called **gz2_class** was introduced to represent the most common consensus morphology classification of galaxies. The **gz2_class** string is generated using the largest debiased volunteer vote fraction beginning from Task 01 and selecting the most probable response through the hierarchical decision tree.

- **Smooth galaxies** are represented by morphology strings beginning with the letter **E**, where:
  - **r** : Completely round galaxy
  - **i** : In-between smooth galaxy
  - **c** : Cigar-shaped smooth galaxy
- **Galaxies with features or disk structures** begin with the letter **S**. Edge-on galaxies are denoted as:
  - **er** : Edge-on galaxy with a round bulge

  - **eb** : Edge-on galaxy with a boxy bulge
  - **en** : Edge-on galaxy without a bulge

- For **oblique disk galaxies**, the presence of a bar is indicated by an uppercase **B**. Bulge prominence is represented as:
  - **d** : No bulge
  - **c** : Just noticeable bulge
  - **b** : Obvious bulge
  - **a** : Dominant bulge

- If a spiral structure is identified, additional symbols describe the winding of the spiral arms:
  - **t** : Tight spiral arms
  - **m** : Medium spiral arms
  - **l** : Loose spiral arms

- Additional unusual galaxy properties may appear at the end of the morphology string:
  - `(r)` : Ring
  - `(l)` : Lens or Arc
  - `(d)` : Disturbed
  - `(i)` : Irregular
  - `(o)` : Other
  - `(m)` : Merger
  - `(u)` : Dust Lane

2.3.1 Galaxy Zoo Challenge Dataset

A subset of the Galaxy Zoo 2 dataset was later released as the Galaxy Zoo Kaggle Challenge dataset. The dataset contains 61,578 labeled RGB galaxy images for training and 79,975 unlabeled images for testing. Each galaxy image is an RGB image of size $424 \times 424 \times 3$ pixels with the target galaxy located at the center of the image.

For each galaxy image, multiple volunteers answered a hierarchical questionnaire consisting of 11 classification questions with 37 possible responses. The final labels are represented as probability distributions obtained from cumulative volunteer voting scores. The challenge objective was to predict the probability distribution of volunteer responses directly from the galaxy images.

The Galaxy Zoo Challenge dataset provides both large-scale image observations and probability-based morphology labels, making it highly suitable for deep learning-based galaxy classification tasks.

2.3.2 High-Confidence Galaxy Zoo 2 Sample Construction

For the present work, the Galaxy Zoo Challenge dataset was synchronized with the Galaxy Zoo 2 classification framework to construct a high-confidence galaxy morphology dataset.

To obtain reliable morphology labels, threshold-based filtering criteria were applied on the cumulative volunteer voting probabilities. Only galaxy images satisfying the predefined confidence thresholds were selected. The threshold conditions used for constructing clean galaxy samples are shown in Table 2.1 [4].

Table 2.1: Threshold Criteria for High-Confidence Galaxy Selection

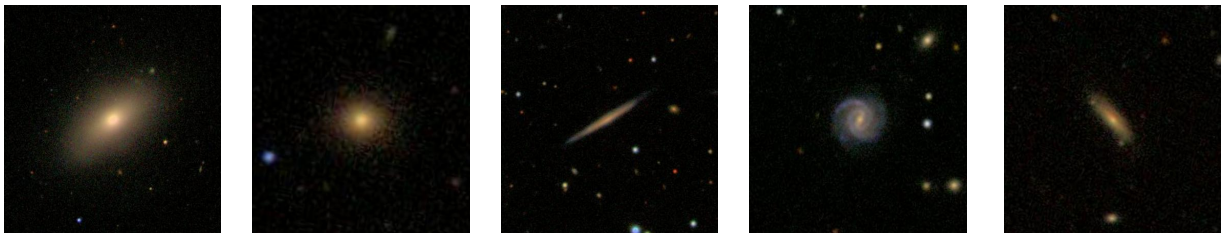
Class	**Task**	**Threshold Condition**
Spiral	T01, T02, T04	$f_{\text{features/disc}} \geq 0.430$, $f_{\text{edge-on,no}} \geq 0.715$, $f_{\text{spiral,yes}} \geq 0.619$
Edge-on	T01, T02	$f_{\text{features/disc}} \geq 0.430$, $f_{\text{edge-on,yes}} \geq 0.602$
Cigar-shaped Smooth	T07, T01	$f_{\text{smooth}} \geq 0.469$, $f_{\text{cigar-shaped}} \geq 0.50$
Completely Round Smooth	T07, T01	$f_{\text{smooth}} \geq 0.469$, $f_{\text{completely round}} \geq 0.50$
In-between Smooth	T07, T01	$f_{\text{smooth}} \geq 0.469$, $f_{\text{in-between}} \geq 0.50$

After applying the threshold filtering rules, 28,790 high-confidence galaxy images were selected for training and evaluation purposes [3].

Table 2.2: Galaxy Zoo 2 dataset distribution after threshold-based filtering

Class ID	Morphology Class	Train	Validation	Total Samples	Proportion (%)
0	In-between smooth	7262	807	8069	28
1	Completely round smooth	7591	843	8434	29
2	Edge-on	3513	390	3903	14
3	Spiral	7025	781	7806	27
4	Cigar-shaped smooth	520	58	578	2
Total		25911	2879	28790	100

Several preprocessing and augmentation techniques were applied to the images, including center cropping, resizing, rotation, and random horizontal and vertical flipping. The original galaxy images of size $424 \times 424 \times 3$ were center cropped and resized before being used in the classification framework.



(a) In-between Smooth (b) Round Smooth (c) Edge-on (d) Spiral (e) Cigar-shaped Smooth

Figure 2.2: High-Confidence Galaxy Zoo 2 samples from the five galaxy morphology classes.

2.4 Galaxy10 DECaLS Dataset

Galaxy10 DECaLS is an improved version of the original Galaxy10 dataset created using Galaxy Zoo annotations combined with higher-quality galaxy images from the DESI Legacy Imaging Surveys [16] (DECaLS). Compared to the original SDSS images, DECaLS images provide significantly improved resolution and image quality.

The dataset contains galaxy images categorized into ten broad morphology classes using volunteer-based annotations and rigorous filtering techniques. The class definitions were refined to improve inter-class separability, while classes containing extremely few samples were removed.

Due to the improved image quality and cleaner class separation, Galaxy10 DECaLS is widely used as a benchmark dataset for deep learning-based galaxy morphology classification. In this work, the dataset is used for additional validation and generalization analysis of the proposed classification framework.

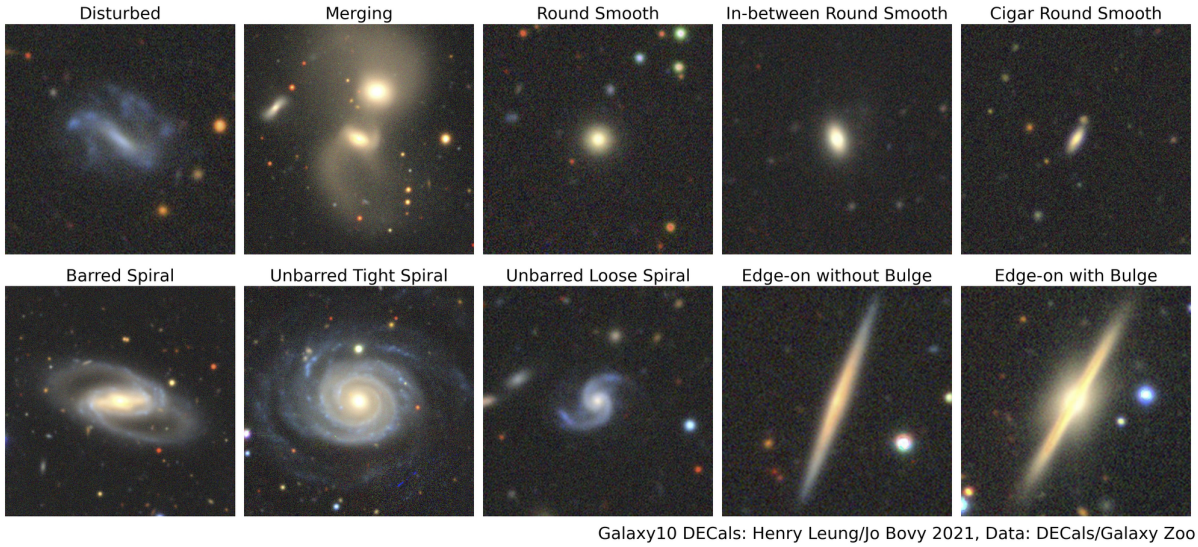


Figure 2.3: Images of different morphology classes from the Galaxy10 DECaLS dataset

2.5 Cross-Dataset Alignment

The objective of this work is to establish a common morphology classification framework between the eCALIFA and Galaxy Zoo 2 datasets for cross-dataset learning (Discussed Earlier). Although both datasets contain galaxy morphology information, their classification schemes are inherently distinct. The eCALIFA dataset provides a single morphology label for each galaxy, whereas Galaxy Zoo 2 employs a hierarchical decision-tree framework that produces a large number of morphology combinations. Therefore, a common taxonomy was required before multimodal training and evaluation could be performed.

2.5.1 Morphology Class Mapping

To obtain a unified classification scheme, Galaxy Zoo 2 morphology labels were mapped into the Galaxy10 DECaLS taxonomy. The original ten-class Galaxy10 DECaLS framework was reduced to eight classes by removing the *Disturbed* and *Merging* categories due to limited correspondence with eCALIFA morphology labels and insufficient sample availability.

To establish a unified classification framework, the mapping rules used in this work are summarized in Table 2.3.

Table 2.3: Mapping of Galaxy Zoo 2 morphology labels to the final eight-class taxonomy

Class ID	**Final Class**	**Galaxy Zoo 2 Mapping Rule**
0	Barred Spiral	Morphology strings beginning with `SB`
1	Cigar-shaped Smooth	`Ec`
2	Edge-on with Bulge	`Ser` and `Seb`
3	Edge-on without Bulge	`Sen`
4	In-between Round Smooth	`Ei`
5	Round Smooth	`Er`
6	Unbarred Loose Spiral	Remaining spiral galaxies containing `L` (loose spiral arms) in the morphology string
7	Unbarred Tight Spiral	Remaining spiral galaxies containing `T` or `M` (tight or medium spiral arms) in the morphology string

This mapping process produced a consistent eight-class morphology taxonomy that could be applied across both datasets.

2.5.2 Galaxy Matching and Dataset Construction

Although Galaxy Zoo 2 morphology labels can be mapped into the Galaxy10 DECaLS taxonomy, the morphology classes of the eCALIFA dataset cannot be directly mapped to either Galaxy Zoo 2 or Galaxy10 DECaLS. This is because eCALIFA follows the traditional Hubble classification system, whereas Galaxy Zoo 2 uses a hierarchical decision-tree framework and Galaxy10 DECaLS employs a different visual morphology taxonomy. Consequently, a direct class-level correspondence between the datasets cannot be established.

To overcome this limitation, galaxy-level matching was performed instead of class-level mapping. Since galaxy identifiers differ across surveys, direct object matching was not feasible. Therefore, galaxy correspondence was established using celestial coordinates

based on Right Ascension (RA) and Declination (DEC), which uniquely identify the position of galaxies on the sky.

Right Ascension (RA) and **Declination (DEC)** [14] are the standard celestial coordinate system used to specify the position of astronomical objects on the sky. Right Ascension is analogous to terrestrial longitude and measures the angular position of an object along the celestial equator, typically expressed in hours, minutes, and seconds. Declination is analogous to latitude and measures the angular distance of an object north or south of the celestial equator in degrees. Together, RA and DEC uniquely identify the location of a galaxy in the celestial sphere, enabling reliable cross-matching of the same object across different astronomical surveys and catalogs.

Using RA–DEC coordinate matching, 231 galaxies common to both eCALIFA and Galaxy Zoo 2 were identified. Since these galaxies were observed in both surveys, the Galaxy Zoo 2 morphology labels could be directly transferred to their corresponding eCALIFA galaxies. The transferred Galaxy Zoo 2 labels were subsequently converted into the Galaxy10 DECaLS eight-class taxonomy using the mapping strategy described in the previous subsection.

After verifying the availability and quality of spectral datacubes, several galaxies containing incomplete observations or corrupted spectral data were removed. This resulted in a final set of 177 valid RGB image–datacube pairs for multimodal training.

The final experimental framework therefore consists of:

- **177 paired RGB images and eCALIFA datacubes for spectral-guided training.**
- **231 matched Galaxy Zoo 2 RGB galaxy images for cross-dataset testing and generalization analysis.**

This strategy enables the transfer of Galaxy Zoo 2 morphology annotations to eCALIFA galaxies through direct galaxy matching while leveraging spectral information during training and evaluating morphology classification performance on an independent galaxy image dataset.

2.6 Triangle Histogram Thresholding for Galaxy Isolation

2.6.1 Application for Galaxy Isolation

In this work, Triangle Histogram Thresholding [12] was employed as a preprocessing step before the final multimodal training and testing stages. The objective was not only to segment the galaxy from the background but also to obtain a more consistent and centered representation of the target galaxy across different observations.

Astronomical RGB images frequently contain background noise, nearby stars, imaging artifacts, and large dark regions that do not contribute to galaxy morphology classification. These unwanted structures can introduce variability during training and reduce the effectiveness of feature extraction. After generating the binary mask, the largest connected component corresponding to the primary galaxy was retained, while smaller foreground objects and background artifacts were removed. The refined mask was then applied to the original RGB image, allowing only the dominant galaxy structure to remain visible.

This preprocessing procedure produced more consistent galaxy-centered images and reduced background interference. As a result, the visual features extracted by the classification network became more representative of the intrinsic galaxy morphology. The centered galaxy representations also improved training stability and enhanced the generalization capability of the final multimodal classification framework.

2.6.2 Thresholding and Segmentation Procedure

Triangle Histogram Thresholding [11] is an adaptive image segmentation technique used to separate foreground objects from the background based on the image intensity distribution. The method is particularly suitable for astronomical images because galaxy observations often contain large dark background regions together with a relatively small bright foreground object.

Initially, each RGB galaxy image is converted into a grayscale representation:

$$I_{gray} = 0.299R + 0.587G + 0.114B$$

where R, G, and B denote the red, green, and blue image channels, respectively.

To suppress high-frequency noise and small intensity fluctuations, Gaussian smoothing is applied:

$$I_{blur}(x, y) = G_\sigma(x, y) * I_{gray}(x, y)$$

where G_σ denotes the Gaussian kernel and $*$ represents convolution.

After smoothing, Triangle Thresholding is applied to determine an adaptive threshold value from the image histogram. The algorithm constructs a line between the histogram peak and the tail of the histogram distribution. The threshold is selected as the point having the maximum perpendicular distance from this line, making the method effective for skewed intensity distributions commonly observed in astronomical images.

The resulting binary segmentation mask is given by

$$M(x, y) = \begin{cases} 1, & I_{blur}(x, y) \geq T \\ 0, & \text{otherwise} \end{cases}$$

where T denotes the threshold obtained using the Triangle Thresholding algorithm.

Connected component analysis is then performed to identify individual foreground regions. The area of each connected component is computed as

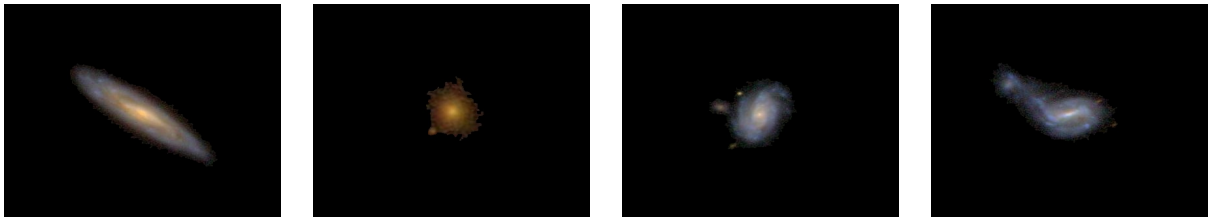
$$A_i = \sum_{x,y} M_i(x, y)$$

where A_i represents the area of the i^{th} connected component. The largest connected component is selected as the primary object of interest. Finally, a morphological closing operation is applied to refine the segmentation mask:

$$M_{closed} = (M \oplus K) \ominus K$$

where $\oplus$ and $\ominus$ denote dilation and erosion operations, respectively, and K represents the structuring kernel.

(a) Original RGB Galaxy Images



(b) Images After Triangle Histogram Thresholding

Figure 2.4: Examples of galaxy images before and after Triangle Histogram Thresholding.

Chapter 3

Related Theory

3.1 Teacher–Student Knowledge Distillation

Knowledge Distillation (KD) is a deep learning framework [6] in which a smaller model, called the *student*, learns knowledge from a larger and more informative model, called the *teacher*. The main objective of knowledge distillation is to transfer the discriminative capability and learned representations of the teacher network into a computationally efficient student network.

Traditional supervised learning trains a model using hard labels represented by one-hot encoded targets. In contrast, knowledge distillation introduces an additional supervision signal using the soft probability distribution generated by the teacher model. These soft predictions contain inter-class similarity information, often referred to as *dark knowledge*, which helps the student learn better feature representations and improved decision boundaries.

The softened class probabilities are generated using temperature-scaled softmax:

$$P_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

where:

- z_i represents the logit corresponding to class i,
- T denotes the temperature parameter,
- P_i represents the softened probability distribution.

A higher temperature produces smoother probability distributions, allowing the student network to learn relationships between different classes rather than relying only on hard labels.

In a Teacher–Student framework, the student model is generally optimized using a combination of supervised classification loss and distillation loss. The overall objective function is commonly expressed as:

$$L_{\text{total}} = \alpha L_{\text{CE}} + (1 - \alpha) L_{\text{KD}}$$

where:

- L_{CE} is the Cross-Entropy loss computed using ground-truth labels,

- L_{KD} represents the knowledge distillation loss,
- α controls the contribution of both loss terms.

The distillation loss is usually computed using Kullback–Leibler (KL) divergence between the teacher and student output distributions:

$$L_{\text{KD}} = KL(P_T^{\text{teacher}} \,||\, P_T^{\text{student}})$$

Knowledge distillation has been widely used for model compression, efficient inference, feature transfer learning, and cross-modal learning tasks. In cross-modal distillation, the teacher network learns from richer or additional modalities, while the student network learns from a simpler modality. This enables the student model to inherit informative representations learned from complex data sources while maintaining low computational cost during inference.

In the proposed framework, knowledge distillation follows a multimodal learning strategy during training. The teacher model learns from two different modalities: RGB galaxy images and spectral representations obtained from datacube features. These two modalities provide complementary information, where RGB images capture spatial morphological structures and spectral data provide wavelength-dependent physical properties of galaxies.

The student model, however, is trained using only RGB images while learning to mimic the predictions of the multimodal teacher network. Through this process, the spectral discriminative knowledge learned by the teacher is transferred into the image-based student model. As a result, the student network can perform efficient galaxy morphology classification using only RGB images during inference while still benefiting from the spectral guidance available during training.

Knowledge distillation improves feature generalization, enhances inter-class separability, and enables efficient deployment of deep learning models without requiring expensive multimodal data during inference.

3.2 Convolutional Vision Transformer (CvT)

The Convolutional Vision Transformer (CvT) is a hybrid architecture [5] that combines convolutional neural networks (CNNs) with Vision Transformers (ViTs). CvT introduces convolutional operations into the Transformer framework to capture both local spatial information and global contextual relationships in images.

Unlike standard Vision Transformers that use fixed image patches, CvT employs convolutional token embedding and convolutional projections inside the self-attention mechanism. This improves local feature extraction while preserving the global learning capability of Transformers.

The self-attention operation is represented as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V denote query, key, and value matrices respectively.

CvT is particularly effective for galaxy morphology classification because it can simultaneously learn fine local structures such as spiral arms and bulges, along with global galaxy morphology patterns.

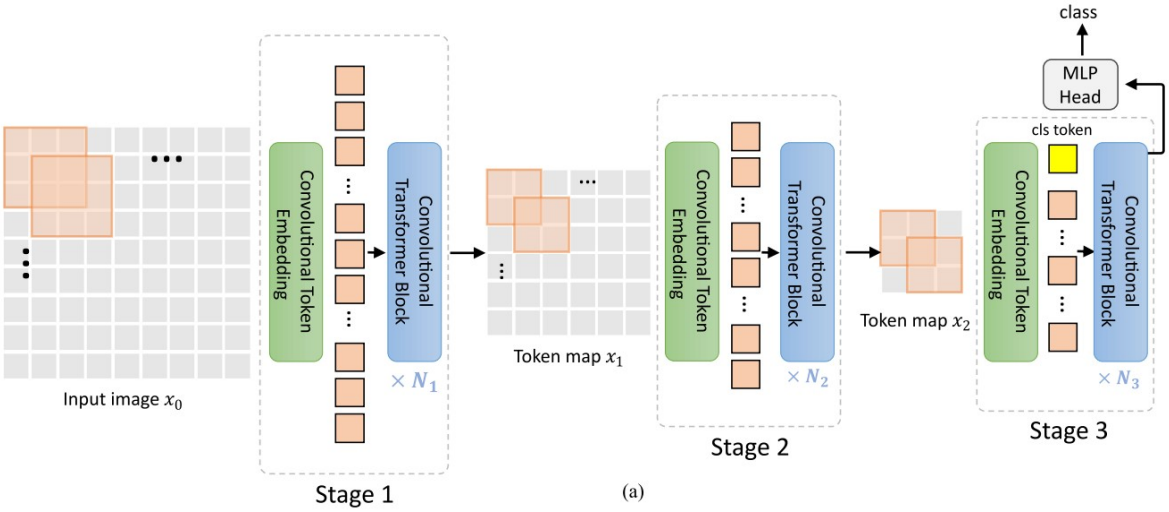


Figure 3.1: Architecture of the Convolutional Vision Transformer (CvT) [5]

In this project, CvT helps learn both fine local galaxy structures and global morphology patterns from RGB galaxy images, making it effective for distinguishing complex galaxy classes. Its combination of convolutional feature extraction and Transformer attention improves morphology-aware representation learning and enhances classification performance.

3.3 Gated Fusion Network

The Gated Fusion Network (GFN) is a multimodal feature fusion framework [10] designed to combine complementary information from different data modalities through adaptive gating mechanisms. In this work, the Gated Fusion framework is used to combine RGB galaxy image features with spectral embedding representations extracted from galaxy datacubes.

The network consists of two parallel feature extraction branches. RGB galaxy images are processed using a pretrained ResNet50 backbone to learn morphology-aware visual representations, while spectral embeddings generated from galaxy datacubes are processed through a multilayer perceptron (MLP) projection network. Both modalities are projected into a common feature space before fusion.

To effectively combine the two modalities, a gated attention mechanism is employed. The RGB and spectral embedding features are concatenated and passed through a gating network that learns adaptive fusion weights. The gating operation is expressed as:

$$G = \sigma\left(W[f_{img}; f_{spec}]\right)$$

where:

- f_{img} denotes RGB image features,
- f_{spec} denotes spectral embedding features,
- $[\cdot;\cdot]$ represents feature concatenation,
- W denotes learnable transformation layers,

- σ represents the sigmoid activation function.

The final fused representation is computed as:

$$f_{fusion} = G \odot f_{img} + (1 - G) \odot f_{spec}$$

where $\odot$ denotes element-wise multiplication.

The gating mechanism enables the network to adaptively regulate the contribution of spectral embeddings based on the visual galaxy features. This allows the model to preserve important morphology-aware visual information while incorporating complementary spectral characteristics from datacube embeddings.

In this work, the fused multimodal representations are used for galaxy morphology classification. The Gated Fusion Network improves cross-modal feature interaction, enhances discriminative representation learning, and enables effective integration of spectral and visual galaxy information.

3.4 Embedding Techniques

3.4.1 Siamese Network

A Siamese Network [1] is a neural network architecture designed to learn similarity-aware feature representations by processing paired inputs through two identical subnetworks with shared parameters. The objective of the Siamese framework is to project semantically similar samples closer together in the embedding space while separating dissimilar samples.

In contrastive learning, Siamese Networks are commonly used to learn invariant feature representations without requiring explicit class supervision. The network learns meaningful embeddings by comparing different augmented views of the same sample. These embeddings preserve important structural and physical characteristics of the data while reducing sensitivity to non-physical variations.

The contrastive learning objective is generally expressed as:

$$L_i = -\log \frac{\exp(\text{sim}(z_i, z_i^+)/\tau)}{\sum_k \exp(\text{sim}(z_i, z_k)/\tau)}$$

where:

- z_i and z_i^+ represent embeddings of two augmented views of the same sample,
- $\text{sim}(\cdot)$ denotes cosine similarity,
- τ is the temperature parameter controlling embedding concentration.

The Siamese architecture learns representations by minimizing the distance between positive pairs while maximizing separation from other samples in the embedding space.

In this work, galaxy datacubes of dimension $192 \times 184 \times 30$ are used to learn spectral embedding representations through the Siamese framework. The Siamese subnetworks use a convolutional neural network (CNN) encoder consisting of multiple convolutional layers with ReLU activation and max-pooling operations for hierarchical feature extraction. The learned embeddings capture important spectral and structural characteristics of galaxies by learning invariant morphology-aware representations from augmented galaxy views.

3.4.2 Dual Shuffle Channel Attention Network

The Dual Shuffle Channel Attention framework is inspired by the Dense Residual Dual-Shuffle Attention Network (DRDA-Net) [9], which combines channel shuffle operations, multi-scale depth-wise convolutions, and channel attention mechanisms for efficient feature learning. In this work, only the Dual Shuffle Residual Block (DRB) and Channel Attention components are utilized to extract compact spectral embedding representations from galaxy datacubes.

The architecture processes galaxy datacubes as spatial-spectral feature maps, where each spectral channel contains wavelength-dependent galaxy information. Initially, a 1×1 convolution operation is applied to obtain compact spectral feature representations.

The extracted feature maps are then passed through the Dual Shuffle Residual Block (DRB). Inside the DRB, the feature maps are divided into two parallel branches and passed through a channel shuffle operation to enable efficient inter-group feature exchange. The outputs processed using 3×3 and 5×5 depth-wise convolutions to capture both local and multi-scale spectral features.Batch normalization followed by another 1×1 convolution operation is then applied, and residual skip connections are incorporated to preserve important spectral information and stabilize feature learning.

Finally, the feature maps are concatenated, refined using convolution operations, and passed through ReLU6 activation for non-linear representation learning. This multi-scale residual architecture enables efficient extraction of discriminative spectral and structural galaxy features with reduced computational complexity.

The channel shuffle operation is defined as:

$$\text{Shuffle}(X) = \text{Reshape} \rightarrow \text{Transpose} \rightarrow \text{Flatten}$$

which facilitates inter-group information exchange between channel groups and improves feature diversity.

To enhance important spectral representations, a Channel Attention mechanism is incorporated. First, global average pooling is applied to aggregate spatial information across channels:

$$F_c = GP(X)$$

where $GP(\cdot)$ denotes the global average pooling operation.

The pooled channel descriptors are then passed through compression and expansion convolution layers followed by activation functions to generate adaptive channel attention weights:

$$M_c(X) = \sigma\left(W_2\left(\delta\left(W_1(\delta(F_c))\right)\right)\right)$$

where:

- W_1 and W_2 denote compression and expansion convolution operations,
- δ represents the ReLU activation function,
- σ denotes the sigmoid activation function.

The generated attention weights are multiplied with the original feature maps to emphasize informative spectral channels while suppressing less relevant features.

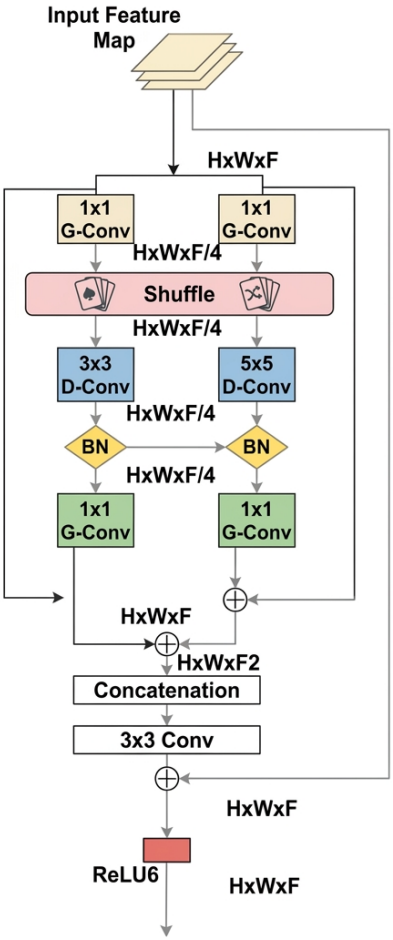


Figure 3.2: Dual Shuffle Residual Block (DRB)

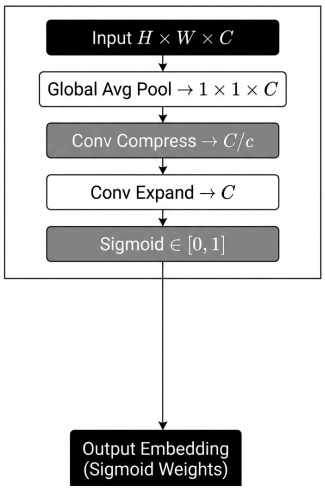


Figure 3.3: Channel Attention Embedding Module

In this work, galaxy datacubes containing spectral channels are resized and processed through the Dual Shuffle Channel Attention framework to learn compact embedding representations. The CNN-based residual architecture, combined with channel shuffle and attention mechanisms, enables efficient extraction of discriminative spectral and structural galaxy features while reducing computational complexity.

3.4.3 Supervised Autoencoder

A Supervised Autoencoder [13] is a neural network architecture that jointly learns latent feature representations and classification-aware embeddings. Unlike a conventional autoencoder, which only reconstructs the input data, a supervised autoencoder incorporates an additional classification objective to guide the latent embedding space toward discriminative feature learning.

In this work, supervised autoencoders are used to generate compact embedding representations from galaxy datacubes. The encoder consists of multiple convolutional layers followed by Batch Normalization and ReLU activation for hierarchical spectral feature extraction. The compressed latent embedding is obtained through a fully connected projection layer, while the decoder reconstructs the original representation using transposed convolution layers. A classification head connected to the latent embedding enables supervised representation learning.

The encoder and decoder operations are defined as:

$$z = f_\theta(x)$$

$$\hat{x} = g_\phi(z)$$

where x denotes the input datacube, z represents the latent embedding, and $\hat{x}$ denotes the reconstructed representation.

The total optimization objective combines reconstruction and classification losses:

$$L_{total} = L_{recon} + L_{CE}$$

where:

$$L_{recon} = \|x - \hat{x}\|^2$$

is the reconstruction loss and L_{CE} denotes the Cross-Entropy classification loss.

The supervised autoencoder learns compact spectral embeddings that preserve important spectral and structural galaxy characteristics while improving discriminative representation learning for downstream analysis.

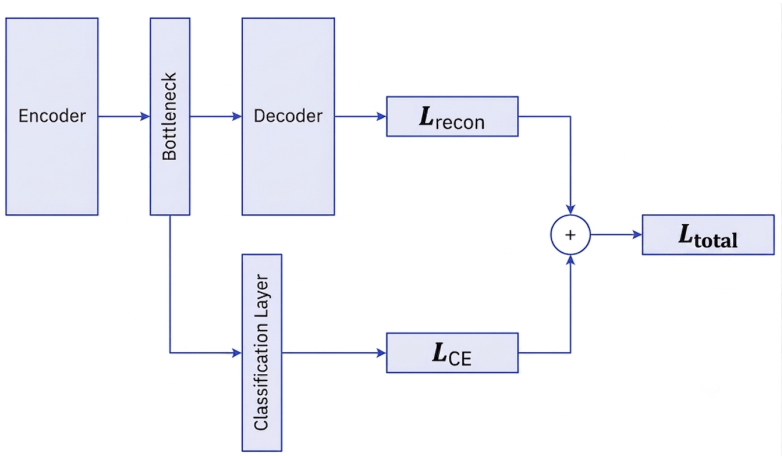


Figure 3.4: Supervised Autoencoder architecture used for spectral embedding generation.

3.5 Embedding Evaluation Metrics

To evaluate the quality and discriminative capability of the learned embedding representations, clustering- and neighborhood-based evaluation metrics are employed. These metrics analyze the compactness, separability, and local consistency of the embedding space. In this work, Silhouette Score, Davies–Bouldin Score, and k-Nearest Neighbor (kNN) evaluation are used to measure the effectiveness of the learned spectral embeddings.

3.5.1 Silhouette Score

The Silhouette Score measures how well samples are clustered with respect to their own class compared to other classes. It evaluates both intra-class compactness and inter-class separation.

For a sample i, the Silhouette coefficient is defined as:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where:

- $a(i)$ represents the average distance between sample i and all other samples within the same class,
- $b(i)$ represents the minimum average distance between sample i and samples belonging to different classes.

The Silhouette Score ranges from -1 to 1:

- Higher values indicate better cluster compactness and class separability,
- Values close to 0 indicate overlapping clusters,
- Negative values indicate incorrect cluster assignments.

A higher Silhouette Score therefore indicates more stable and discriminative embedding representations.

3.5.2 Davies–Bouldin Score

The Davies–Bouldin (DB) Score evaluates cluster similarity by measuring the ratio between intra-cluster dispersion and inter-cluster separation.

The DB score is defined as:

$$DB = \frac{1}{N} \sum_{i=1}^{N} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right)$$

where:

- S_i and S_j denote intra-cluster distances for clusters i and j,
- M_{ij} represents the distance between cluster centroids,

- N denotes the total number of clusters.

Lower Davies–Bouldin values indicate:

- smaller intra-class variation,
- better inter-class separation,
- more compact and stable embedding distributions.

Higher DB scores indicate overlapping clusters and reduced embedding discriminability.

3.5.3 k-Nearest Neighbor (kNN) Evaluation

The k-Nearest Neighbor (kNN) evaluation measures local neighborhood consistency within the embedding space. For each embedding vector, the k closest neighboring samples are identified using distance-based similarity measures such as Euclidean distance or cosine similarity.

The Euclidean distance between two embedding vectors is defined as:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_i^k - x_j^k)^2}$$

where x_i and x_j denote embedding vectors.

A higher kNN classification accuracy indicates that samples belonging to the same morphology class remain close together in the embedding space. This reflects improved local feature consistency and stronger representation stability.

Together, these metrics provide complementary evaluation of the learned embedding space by analyzing cluster compactness, inter-class separability, and neighborhood-level consistency of the spectral representations.

Chapter 4

Methodology

4.1 Preliminary Experiments

4.1.1 PCA-Based Spectral Classification

The eCALIFA spectral datacubes (Section 2.1) were initially compressed using Principal Component Analysis (PCA) to obtain compact spectral representations suitable for morphology classification [7].

The cumulative explained variance was analyzed across the principal components. It was observed that after the first 12 components, the increase in explained variance became marginal. Therefore, only the first 12 PCA components were retained for further experiments.

A total of 895 eCALIFA datacubes were processed, where each galaxy was represented using a 12-dimensional PCA feature vector.The dataset was divided using an 80:20 train-test split strategy. The PCA feature vectors were then provided to a Random Forest classifier [8] for morphology classification.

The classifier achieved approximately **92–93%** testing accuracy, indicating that the PCA-compressed spectral representations preserve significant morphology-aware information from the original datacubes.

4.1.2 CNN-Based RGB Image Classification

The corresponding RGB galaxy images (Section 2.2) were collected from SDSS DR10 observations. Due to corruption, noise, and mismatched samples, only 786 RGB galaxy images were retained after filtering. The original eCALIFA morphology labels were preserved in their 21-class form to evaluate the feasibility of direct RGB-based galaxy morphology classification.

Initially, a standard CNN architecture was trained directly on the RGB images using an 80:20 train–test split. However, the model produced poor classification performance, achieving a testing accuracy of only **14.50%**.

The low performance was primarily attributed to poor image quality, observational noise, class imbalance, and the difficulty of distinguishing fine-grained morphology classes using RGB information alone.

4.2 Spectral Knowledge Transfer using Teacher–Student Knowledge Distillation

The eCALIFA dataset originally contains 21 morphology classes based on the Hubble classification system. To reduce class imbalance and obtain a more robust classification framework, these classes were consolidated into five broader categories: BCD (Blue Compact Dwarf), E (Elliptical; E0–E7), I (Irregular), S (Spiral/Lenticular; S0–Sdm), and U (Uncertain/Unclassified). This consolidation increases the number of samples within each class while preserving the major structural characteristics required for morphology classification.

To improve RGB image classification performance, a Teacher–Student knowledge distillation framework was implemented based on the concepts (Discussed in Section 3.1). The objective was to transfer morphology-related information learned from spectral observations into an RGB image-based classification model.

4.2.1 Proposed Cross-Modal Distillation Framework

- Traditional Teacher–Student frameworks generally operate on a single modality. In this work, a cross-modal distillation strategy was introduced where the teacher learns jointly from spectral and visual information while the student receives only RGB images.
- The inclusion of spectral features allows the teacher to learn wavelength-dependent galaxy characteristics that are not directly visible in RGB observations, resulting in richer morphology-aware representations.
- Through knowledge distillation, this spectral information is transferred to the RGB-only student model, improving classification performance without requiring spectral datacubes during testing.

4.2.2 Teacher–Student Architecture

- The teacher network receives an RGB galaxy image of size $3 \times 128 \times 128$ and a 12-dimensional PCA feature vector.
- The RGB branch in Teacher network consists of convolutional layers ($3 \rightarrow 32 \rightarrow 64$) followed by global average pooling and a fully connected layer to produce a 128-dimensional feature representation.
- The PCA branch in Teacher Network uses an MLP ($12 \rightarrow 128 \rightarrow 128$) to generate a 128-dimensional spectral representation.
- The image and spectral features are concatenated to form a 256-dimensional multimodal feature vector, which is passed through a classifier ($256 \rightarrow 128 \rightarrow N_c$).
- The teacher network is trained using Cross-Entropy loss on the morphology labels.
- The student network contains only the RGB branch and a classifier ($128 \rightarrow 64 \rightarrow N_c$). It is trained using both Cross-Entropy loss and knowledge distillation from the pretrained teacher.
- During inference, only RGB images are required by the student network.

4.2.3 Training Configuration

The paired RGB dataset containing 786 samples was divided using an 80:20 train–test split strategy.

The teacher and student networks were trained using the Adam optimizer with a learning rate of 1×10^{-3}. Cross-Entropy loss and KL-Divergence loss were jointly used during optimization. The temperature parameter for distillation was set to $T = 3.0$, while the balancing parameter was fixed at $\alpha = 0.5$.

Both teacher and student models were trained using a batch size of 8. The teacher network was trained for 50 epochs, followed by student distillation training for another 50 epochs.

4.2.4 Classification Performance

After knowledge distillation, the RGB-only student model achieved a classification accuracy of **98.55%**. The results indicate that the multimodal teacher successfully transferred morphology-discriminative spectral information to the student network. Consequently, the student model achieved high classification performance while requiring only RGB galaxy images during inference. The detailed classification report is summarized in Table 4.1.

Table 4.1: Classification performance of the distilled student model.

Class	Precision	Recall	F1-Score	Support
BCD	1.00	0.33	0.50	878
E	1.00	1.00	1.00	13931
I	1.00	0.70	0.83	1384
S	0.98	1.00	0.99	59398
U	1.00	0.67	0.80	309
Accuracy	98.55%			
Macro Avg	1.00	0.74	0.82	75900
Weighted Avg	0.99	0.99	0.98	75900

4.3 Galaxy Zoo 2 Morphology Classification using CvT

The high-confidence Galaxy Zoo 2 dataset (Described in Section 2.3.2) was used for RGB galaxy morphology classification. The final filtered dataset contains 28,790 RGB galaxy images belonging to five morphology classes obtained using the probability threshold criteria previously discussed.

Only rotational augmentation was applied during training to improve robustness against orientation variation in galaxy structures. The minority class corresponding to cigar-shaped smooth galaxies was additionally balanced using rotational augmentation.

4.3.1 Architecture and Training Configuration

For morphology classification, a Convolutional Vision Transformer (CvT) architecture was implemented (Discussed in Section 3.2). The model combines convolutional token embedding with Transformer-based self-attention to learn both local galaxy structures and global morphology-aware representations.

The input RGB galaxy images were resized to 224×224 resolution before training.

The CvT architecture consists of three hierarchical stages:

- Stage 1: $3 \rightarrow 64$ channels, 1 Transformer block, 1 attention head
- Stage 2: $64 \rightarrow 128$ channels, 2 Transformer blocks, 2 attention heads
- Stage 3: $128 \rightarrow 256$ channels, 10 Transformer blocks, 4 attention heads

The dimensional progression through the network is:

- Input tensor: $(B, 3, 224, 224)$
- Stage 1 feature map: (56×56)
- Stage 2 feature map: (28×28)
- Stage 3 feature map: (14×14)

After the final Transformer stage, global average pooling was applied to obtain a 256-dimensional feature representation, which was passed through a fully connected classification layer for five-class galaxy morphology prediction.

The network was trained using the Adam optimizer with a learning rate of 1×10^{-5} and Cross-Entropy loss for 150 epochs.

4.3.2 Classification Performance

The CvT model achieved a validation accuracy of **85.84%**. The detailed class-wise classification performance is summarized in Table 4.2.

Table 4.2: Classification report for the CvT model

Class	Precision	Recall	F1-score	Support
Cigar-shaped smooth	0.37	0.47	0.42	116
Completely round	0.88	0.91	0.90	1688
Edge-on	0.88	0.85	0.87	781
In-between	0.85	0.84	0.84	1614
Spiral	0.88	0.85	0.87	1562
Accuracy	0.86			
Macro Average	0.77	0.79	0.78	5761
Weighted Average	0.86	0.86	0.86	5761

4.4 Cross-Dataset Evaluation using PCA Features

After evaluating the proposed Teacher–Student Knowledge Distillation framework on the 177 paired eCALIFA RGB image–datacube dataset and the CvT-based image classification model on the Galaxy Zoo 2 dataset, both approaches achieved satisfactory classification performance on their respective datasets. However, the primary objective of this work was not limited to within-dataset classification. Instead, the goal was to investigate whether spectral information learned from eCALIFA datacubes could improve galaxy morphology classification on an independent dataset containing only RGB galaxy images.

To evaluate this capability, a cross-dataset experiment was conducted using the Teacher–Student framework described in Section 4.2. The teacher model was trained using paired eCALIFA RGB images and PCA-compressed spectral features, while testing was performed on 231 Galaxy Zoo 2 images(Section 2.5.2) after converting their morphology labels into the same eight-class taxonomy. Apart from replacing the testing dataset, the overall architecture and training procedure remained unchanged.

A significant performance degradation was observed under this cross-dataset setting, with the classification accuracy dropping to approximately **5%–8%**. This result suggested that the PCA-compressed spectral representations were unable to preserve sufficient morphology-discriminative information for reliable transfer across datasets.

To further investigate whether the limitation originated from the fusion strategy rather than the PCA features themselves, the multimodal architecture was subsequently modified using a stronger feature fusion framework.

Improved Multimodal Fusion using PCA Features

To investigate whether the poor cross-dataset performance originated from the fusion strategy rather than the PCA features themselves, the Teacher–Student framework was replaced with a multimodal fusion architecture inspired by bilinear fusion networks.

- A pretrained ResNet34 backbone was used to extract a 512-dimensional feature vector from the RGB galaxy image.
- The 12-dimensional PCA feature vector was projected through an MLP ($12 \rightarrow 64 \rightarrow 256$) to obtain a higher-dimensional spectral representation.

- The image and spectral representations were fused through feature concatenation, producing a 768-dimensional multimodal feature vector (512 + 256).
- The fused representation was passed through a classification network (768 → 512 → 8) to predict the galaxy morphology class.
- Random PCA feature dropout was introduced during training to improve robustness and reduce over-reliance on spectral features.
- During testing, only matched RGB galaxy images from the Galaxy Zoo 2 dataset (Discussed in section 2.5.2) were provided to the model.

The modified fusion architecture improved the cross-dataset classification accuracy from approximately 5–8% to **14.49%**.

Table 4.3: Cross-dataset performance of the PCA-based multimodal fusion framework.

Accuracy	Macro Precision	Macro Recall	Macro F1-score
14.49%	32.36%	17.44%	14.69%

In addition to the improvement in classification performance, the proposed fusion framework provides a simpler training strategy compared to the Teacher–Student architecture, as it eliminates the need for separate teacher and student networks and avoids the additional optimization overhead associated with knowledge distillation. **Furthermore, the multi-stage optimization procedure of the Teacher–Student framework tended to learn dataset-specific characteristics from the training data, resulting in poor generalization when the testing data originated from a different distribution. In contrast, the fusion-based framework exhibited better cross-dataset robustness by learning a shared multimodal representation directly.**

Despite these advantages, the overall performance remained insufficient for reliable morphology classification. This suggests that the primary limitation originated from the discriminative quality of the PCA representations rather than the fusion architecture itself, motivating the exploration of more informative spectral embedding techniques.

4.5 Motivation for Embedding-Based Spectral Representations

The results obtained using PCA-based spectral features indicate that the primary limitation lies in the quality of the spectral representations rather than the multimodal fusion architecture. To obtain more discriminative and morphology-aware spectral features, embedding-based representation learning techniques were subsequently investigated. Unlike PCA, embedding networks learn feature representations directly from the data and can capture complex morphology-related spectral patterns.

4.5.1 Siamese Embedding and Outcomes

The Siamese Network framework (Discussed in Section 3.4.1) was employed to generate 512-dimensional spectral embeddings from the eCALIFA datacubes.

During Siamese network training, contrastive learning was performed using pairs of augmented datacubes. The augmentation strategy included:

- Gaussian noise injection to simulate different signal-to-noise observational conditions.
- Random rotations and horizontal/vertical flips to improve orientation invariance.
- Gaussian blurring to mimic observational and instrumental degradation effects.
- Random spatial translations to reduce sensitivity to the exact galaxy position within the field of view.

These augmentations encouraged the network to learn robust and invariant spectral representations while preserving morphology-related spectral characteristics.

After training, the generated 512-dimensional Siamese embeddings were integrated into the multimodal fusion framework (Described in the section 4.4) by replacing the PCA representations. Compared to the PCA-based approaches, the Siamese embeddings produced substantially more discriminative spectral features and improved cross-dataset generalization.

The resulting model achieved a classification accuracy of **46.75%** on the Galaxy Zoo 2 test dataset. This significant improvement indicates that the Siamese network was able to capture morphology-relevant spectral information more effectively than PCA-based representations. The confusion matrix obtained using the Siamese embeddings is shown in Figure 4.1.

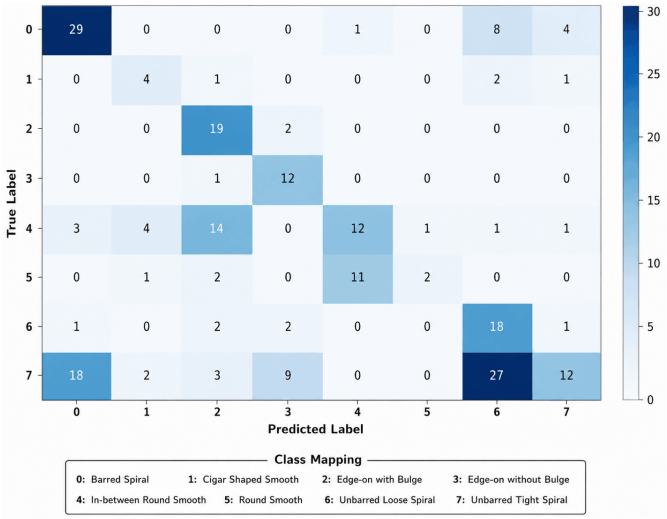
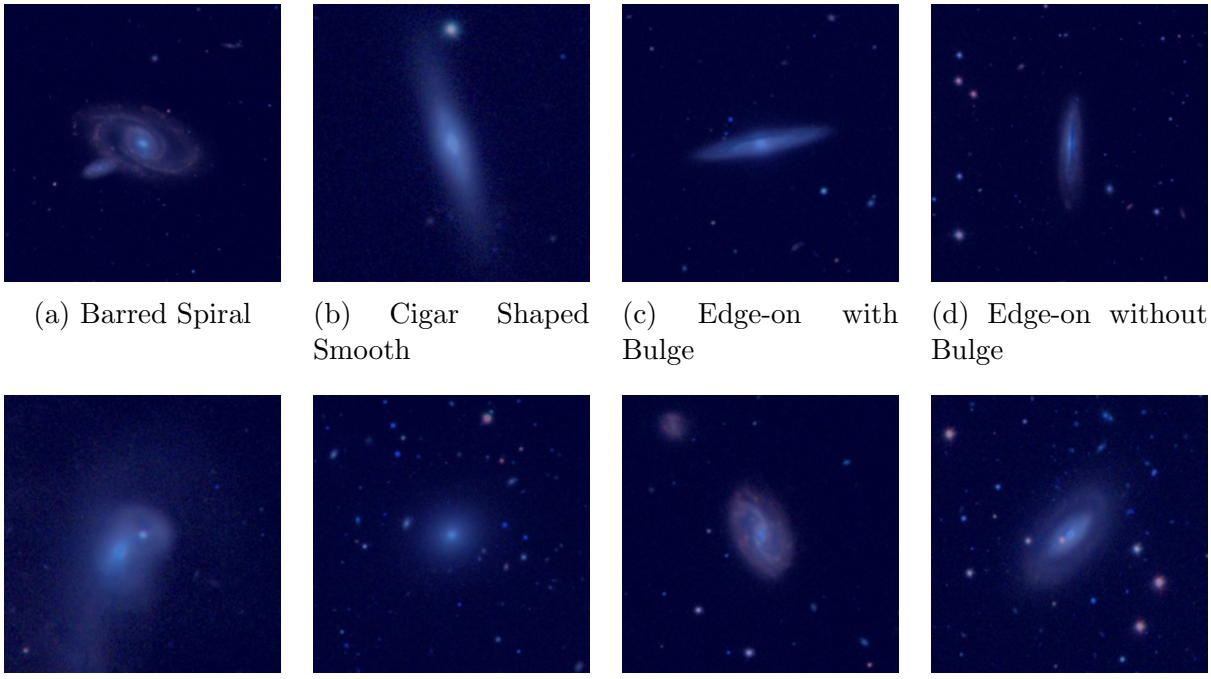


Figure 4.1: Confusion matrix obtained using Siamese embeddings within the multimodal fusion framework.

The Grad-CAM visualizations show that the model generally focuses on the galaxy itself, but the highlighted regions are often broad and not well localized. Features such as bars, spiral arms, bulges, and edge-on disks are not consistently emphasized, and in some cases the activations extend beyond the visible boundaries of the galaxy. These observations are consistent with the confusion matrix results, where substantial confusion exists between morphologically similar galaxy classes. The weak localization of the activation maps suggests that the spectral embeddings provide only limited guidance toward morphology-relevant structures during multimodal learning.



(a) Barred Spiral (b) Cigar Shaped Smooth (c) Edge-on with Bulge (d) Edge-on without Bulge

(e) In-between Round Smooth (f) Round Smooth (g) Unbarred Loose Spiral (h) Unbarred Tight Spiral

Figure 4.2: Examples of class-wise activation heatmaps generated from the multimodal classification network trained using Siamese spectral embeddings. The highlighted regions indicate the image areas that contributed most strongly to the model's predictions for each galaxy morphology class.

Analysis of Siamese Embedding Limitations

Although the Siamese embeddings improved the classification accuracy compared to the PCA-based representations, the overall performance was still limited. This can be attributed to the objective of Siamese learning, which focuses on capturing similarity relationships between samples rather than separating galaxies according to their morphology classes. As a result, galaxies from different classes can still occupy nearby regions in the embedding space, leading to overlap between categories. This behaviour is evident from both the confusion matrix and the Grad-CAM visualizations. While the embeddings successfully capture spectral similarity information, they do not consistently encode morphology-specific characteristics, motivating the exploration of embedding methods that incorporate class supervision directly during training.

4.5.2 Proposed Dual Shuffle Channel Attention Embedding Technique

Unlike the Siamese Network, which learns a generalized similarity-aware embedding space, the proposed Dual Shuffle Channel Attention (DCA) framework is explicitly guided by the galaxy morphology classes during training. Consequently, the generated embeddings are optimized not only to capture spectral characteristics but also to improve class-level separability in the embedding space.

The Dual Shuffle Channel Attention (DCA) embedding framework (Discussed in Section 3.4.2) was inspired by the DRDA-Net architecture. While the original DRDA-Net was designed as a complete image classification network, the objective of this work was to generate compact morphology-aware spectral embeddings from eCALIFA datacubes for downstream multimodal galaxy morphology classification.

To adapt the architecture for spectral representation learning, the following modifications were introduced:

- The original DRDA-Net was designed for direct image classification. Only the feature extraction and channel-attention components were retained.
- The final classification layer was retained only during training to provide supervision for feature learning. The 512-dimensional feature representation was passed through a fully connected classification layer to predict the eight galaxy morphology classes.
- Cross-Entropy loss was computed between the predicted morphology labels and the ground-truth galaxy classes. During backpropagation, this loss updated the channel-attention mechanism, enabling it to assign higher importance to morphology-relevant spectral channels.
- After training, the classification output was discarded and the learned 512-dimensional channel-attention feature vector was extracted as the final spectral embedding representation.

These modifications transformed the original classification architecture into a lightweight morphology-aware spectral embedding generator. By directly incorporating morphology supervision during training, the generated embeddings emphasize discriminative spectral channels while suppressing less informative features, making them more suitable for downstream multimodal galaxy morphology classification than generalized similarity-based embeddings.

4.5.3 Supervised Autoencoder and Input Data Adaptation

Although the proposed Dual Shuffle Channel Attention framework was able to generate morphology-aware spectral embeddings from the complete eCALIFA datacubes, the extremely high dimensionality of the spectral observations introduces considerable computational and memory requirements. To investigate whether compact yet discriminative spectral representations could be learned from a reduced spectral region, a supervised autoencoder framework was also explored.

While the general encoder–decoder architecture follows the supervised autoencoder concept (Discussed in Section 3.4.3), modifications were introduced at the input level to adapt the model to the high-dimensional eCALIFA datacubes.

The following modifications were introduced to adapt the supervised autoencoder for eCALIFA datacubes:

- **Input Adaptation:** The original eCALIFA datacubes contain 1877 spectral channels, making direct processing computationally expensive. Analysis of the spectral distribution indicated that a large fraction of morphology-relevant information is concentrated around the central wavelength region. Therefore, only the central 256 spectral channels were retained and resized to a fixed spatial resolution before being provided to the network.

- **Architecture:** The network consists of a convolutional encoder that compresses the selected 256-channel spectral cube into a 512-dimensional latent representation, a decoder that reconstructs the input cube from the latent space, and an additional eight-class morphology classification head connected directly to the latent embedding layer.

The reconstruction objective encourages the latent representation to preserve spectral information, while the classification objective forces the embedding space to organize according to galaxy morphology classes. As a result, the generated 512-dimensional embeddings become highly morphology-aware and class-discriminative, making them suitable for downstream multimodal learning.

4.5.4 Comparative Analysis of Embedding Quality

Before integrating the generated embeddings into the final multimodal classification framework, it is important to evaluate whether the learned embedding space possesses sufficient discriminative capability for morphology classification. Training a multimodal classifier using poor-quality embeddings can lead to suboptimal performance regardless of the fusion architecture employed. Therefore, the generated embeddings were first assessed using clustering- and neighborhood-based evaluation metrics to measure their compactness, separability, and class-discriminative properties.

The embedding quality was evaluated using the Silhouette Score, Davies–Bouldin Score, and k-Nearest Neighbor (kNN) classification accuracy (Discussed in section 3.5). These metrics provide complementary insights into the structure of the learned embedding space and help determine whether the embeddings are suitable for subsequent multimodal learning.

Table 4.4: Comparative evaluation of different embedding techniques

Embedding Method	Silhouette Score	Davies–Bouldin Score	kNN Accuracy
1. Siamese Network	-0.1794	5.7804	0.2603
2. Dual Shuffle Channel Attention Network	-0.0970	3.5843	0.5919
3. Supervised Autoencoder	0.2900	1.2232	0.9944

The evaluation results indicate clear differences in the quality of the learned spectral representations. The Siamese network produced generalized similarity-aware embeddings but exhibited poor cluster separation, resulting in relatively low kNN classification performance. The Dual Shuffle Channel Attention framework improved the structure of the embedding space by emphasizing morphology-relevant spectral channels, leading to better clustering and neighborhood consistency. Among all the investigated methods, the Supervised Autoencoder generated the most discriminative embeddings, achieving the highest Silhouette Score, the lowest Davies–Bouldin Score, and the best kNN classification accuracy. These results suggest that incorporating morphology supervision during embedding learning substantially improves the quality of the learned spectral representations.

4.6 Multimodal Classification using Gated Fusion Network

Although the Teacher–Student framework and the multimodal fusion architecture for cross-dataset galaxy morphology classification demonstrated the usefulness of spectral information, they also exhibited certain limitations. The Teacher–Student framework required a relatively complex two-stage training procedure, while the multimodal fusion model relied primarily on direct feature concatenation to combine image and spectral information.

To overcome these limitations, a Gated Fusion Network (GFN) (Described in Section 3.3) was investigated. Unlike simple concatenation-based fusion, the GFN employs a learnable gating mechanism that dynamically determines the contribution of each modality during feature fusion. This enables the network to emphasize the most informative modality for a given sample and combine image and spectral representations in a more effective manner. Consequently, the proposed framework provides a more flexible and adaptive approach for multimodal galaxy morphology classification. In addition, to improve the localization of galaxy structures and reduce background activations observed in the Siamese Grad-CAM visualizations, the triangle histogram thresholding method (Described in Section 2.6) was applied during both training and testing for image centering and foreground extraction.

4.6.1 Input Preprocessing

The framework was trained using RGB galaxy images collected from SDSS DR10 together with spectral embeddings generated from the corresponding eCALIFA datacubes (177 pairs). Prior to training, all RGB images were resized to a fixed resolution of 224×224 pixels and converted into tensor representations. Channel-wise normalization was then

applied using the ImageNet mean and standard deviation values:

$$x' = \frac{x - \mu}{\sigma}$$

where x denotes the original pixel intensity, μ represents the channel mean, and σ denotes the corresponding channel standard deviation. The normalization parameters used were $\mu = (0.485, 0.456, 0.406)$ and $\sigma = (0.229, 0.224, 0.225)$ for the red, green, and blue channels, respectively. Image resizing ensures a uniform input size for the ResNet50 backbone, while normalization stabilizes feature distributions across batches and facilitates efficient optimization during training.

4.6.2 Architecture Overview

The proposed multimodal framework consists of two parallel branches:

- RGB Galaxy Image Branch
- Spectral Embedding Branch

RGB Feature Extraction Branch

For visual feature extraction, a pretrained ResNet50 backbone initialized with ImageNet weights was employed. RGB galaxy images of size $224 \times 224 \times 3$ were provided as input to the network. After the final convolutional stage and global average pooling, ResNet50 produced a 2048-dimensional feature vector. This representation was subsequently projected through fully connected layers from $2048 \rightarrow 512 \rightarrow 256$, with Batch Normalization, ReLU activation, and Dropout regularization applied between layers. The resulting 256-dimensional latent embedding served as the morphology-aware RGB feature representation used for multimodal fusion.

Spectral Embedding Branch

The spectral branch processes the pre-computed 512-dimensional embeddings generated from eCALIFA datacubes using a multilayer perceptron (MLP). The embeddings are transformed through fully connected layers with dimensions $512 \rightarrow 512 \rightarrow 256$, together with Batch Normalization, ReLU activation, and Dropout regularization. This projection maps the spectral information into the same 256-dimensional latent space as the RGB branch, enabling effective feature alignment and multimodal fusion.

Final Classification Layer

The fused 256-dimensional multimodal representation is passed through fully connected classification layers with dimensions $256 \rightarrow 128 \rightarrow 8$, together with ReLU activation and Dropout regularization. The final output layer produces the probability scores for the eight galaxy morphology classes.

4.6.3 Proposed Modifications to the Gated Fusion Network

To combine information from both modalities, the 256-dimensional RGB feature vector and the 256-dimensional spectral feature vector are concatenated to form a 512-dimensional representation. This concatenated feature vector is passed through a sigmoid-based gating network with dimensions $512 \rightarrow 256$ to generate adaptive gate weights.

Unlike the original GFNet formulation, a modified fusion strategy was introduced in this work to better preserve morphology-aware visual information while selectively incorporating spectral knowledge. The fused representation is computed as

$$F = F_{RGB} + G \odot F_{Spec}$$

where F_{RGB} and F_{Spec} denote the RGB and spectral feature representations, respectively, G represents the learned gate weights, and $\odot$ denotes element-wise multiplication. **In this formulation, the RGB feature representation acts as the primary morphology descriptor, while the spectral features contribute through the learned gating mechanism**.

This allows the network to dynamically regulate the influence of spectral information for each galaxy sample, emphasizing informative spectral characteristics while suppressing less relevant features. The resulting fused representation remains 256-dimensional and is subsequently used for final morphology classification.

4.6.4 Integrating Auxiliary Spectral Supervision

To further improve feature learning, an auxiliary classifier was incorporated into the spectral branch. The 256-dimensional spectral embedding is directly connected to an auxiliary classification layer for morphology prediction. This additional supervision encourages the spectral embeddings to retain morphology-discriminative information even before multimodal fusion takes place.

4.6.5 Optimization Strategy

The network was optimized using the Adam optimizer with a learning rate of 3×10^{-4} and weight decay regularization. Adam was selected because of its adaptive learning-rate capability and stable convergence behavior for multimodal deep learning architectures. A OneCycle learning rate scheduling strategy was employed to accelerate convergence during the initial training stages while enabling fine-grained optimization near convergence.

4.6.6 Proposed Loss Function Design

The original GFNet architecture employs a single optimization objective during training. In the present work, the training strategy was modified to better support multimodal galaxy morphology classification. Instead of relying on a single loss function, multiple complementary objectives were jointly optimized to improve morphology classification performance, preserve discriminative spectral information, and strengthen the alignment between RGB and spectral feature representations.

- **Classification Loss (L_{cls})**

  The primary objective was the Cross-Entropy classification loss:

$$L_{cls} = CE(y, \hat{y})$$

  where y denotes the ground-truth morphology label and $\hat{y}$ represents the predicted class probabilities.

- **Auxiliary Classification Loss (L_{aux})**

  An additional Cross-Entropy loss was applied to the spectral embedding branch through an auxiliary classifier. This encourages the spectral embeddings to preserve morphology-discriminative information before multimodal fusion.

- **Cosine Alignment Loss (L_{align})**

  To align RGB and spectral representations within a common latent space, cosine similarity loss was employed:

$$L_{align} = 1 - \frac{F_{RGB} \cdot F_{Spec}}{\|F_{RGB}\| \|F_{Spec}\|}$$

  where F_{RGB} and F_{Spec} denote the RGB and spectral feature representations, respectively. This loss minimizes the angular distance between modalities and improves feature consistency.

- **Contrastive Loss ($L_{contrast}$)**

  To further enhance cross-modal discriminability, a contrastive loss based on normalized feature similarity was introduced:

$$L_{contrast} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(\frac{F_{RGB}^{(i)} \cdot F_{Spec}^{(i)}}{\tau}\right)}{\sum_{j=1}^{N} \exp\left(\frac{F_{RGB}^{(i)} \cdot F_{Spec}^{(j)}}{\tau}\right)}$$

  where τ denotes the temperature parameter and N represents the batch size. This objective pulls matching RGB–spectral pairs closer together while pushing unrelated pairs apart within the shared embedding space.

The overall training objective is given by:

$$L_{total} = L_{cls} + 0.5L_{aux} + 0.3L_{align} + 0.3L_{contrast}$$

The weighting coefficients were selected empirically to prioritize morphology classification performance while simultaneously enforcing feature alignment and cross-modal consistency. The model was trained for 150 epochs using a batch size of 16.

4.6.7 Results using Supervised Autoencoder Embeddings

The spectral embeddings generated using the Supervised Autoencoder framework were integrated with RGB galaxy images through the proposed Gated Fusion Network. Since the embeddings were learned under morphology supervision, the latent representations preserved class-discriminative spectral information.

The multimodal fusion of RGB and spectral features achieved an overall classification accuracy of **73.59%** on test set. The corresponding confusion matrix is shown in Figure 4.3.

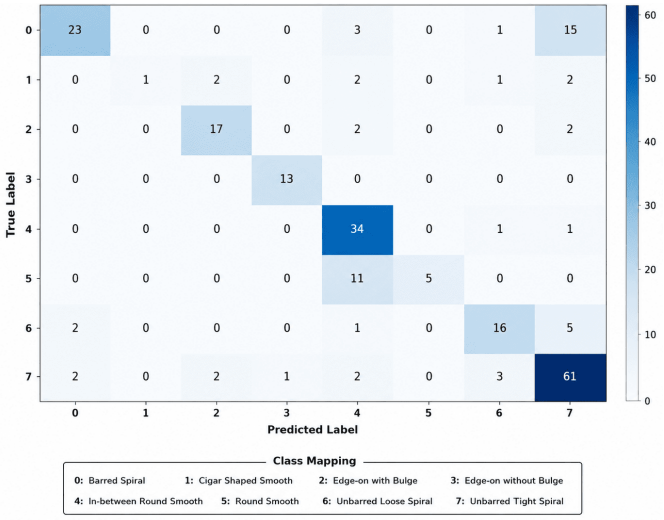


Figure 4.3: Confusion matrix obtained using Supervised Autoencoder embeddings within the Gated Fusion Network framework.

The results indicate that the supervised spectral representations improve morphology classification performance by incorporating class-aware spectral information. However, noticeable confusion remains between visually similar smooth and spiral subclasses.

4.6.8 Results using Dual Shuffle Channel Attention Embeddings

The Dual Shuffle Channel Attention framework was subsequently evaluated within the same Gated Fusion architecture. Unlike the Supervised Autoencoder, these embeddings were generated through contrastive spectral feature learning and channel-attention-based representation extraction.

The resulting multimodal classifier achieved a best test accuracy of **77.15%**. The corresponding confusion matrix is shown in Figure 4.4.

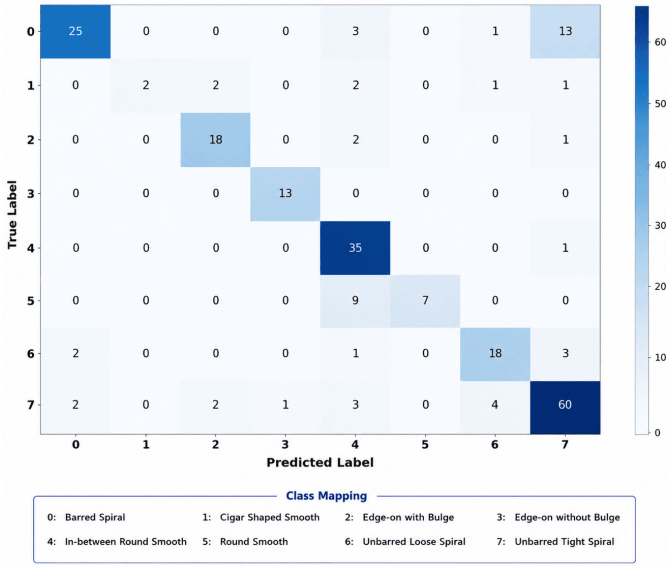


Figure 4.4: Confusion matrix obtained using Dual Shuffle Channel Attention embeddings within the Gated Fusion Network framework.

Compared to the Supervised Autoencoder embeddings, the Dual Shuffle Channel Attention embeddings produced better class separation and improved overall classification performance. The learned spectral representations were more effective at preserving discriminative galaxy characteristics, resulting in reduced confusion among several morphology classes and improved multimodal fusion.

4.6.9 Comparative Analysis

Among the investigated embedding techniques, the Dual Shuffle Channel Attention embeddings achieved the best overall performance within the Gated Fusion framework. A possible reason is that the attention-based model utilizes the complete spectral datacube during embedding generation, whereas the Supervised Autoencoder operates on the selected central spectral channels. While the reduced input makes the autoencoder computationally efficient, some useful spectral information may be lost in the process. Nevertheless, both approaches produced a substantial improvement over the earlier PCA-based and Siamese-based representations, highlighting the importance of morphology-aware embeddings and adaptive multimodal fusion for cross-dataset galaxy morphology classification.

Chapter 5

Conclusion and Future Work

This project explored the use of multimodal learning for galaxy morphology classification by combining spectral information from eCALIFA datacubes with RGB galaxy images. Different spectral embedding techniques, including Siamese Networks, Dual Shuffle Channel Attention, and Supervised Autoencoders, were investigated to understand how effectively spectral information can support morphology classification. The generated embeddings were evaluated using both Teacher–Student and Gated Fusion architectures. The results showed that incorporating spectral information alongside RGB images improves classification performance compared to relying on visual information alone. Among the approaches explored, the Gated Fusion framework provided the most effective integration of spectral and visual features and achieved the best overall performance. However, the study was constrained by the limited number of matched galaxies available after RA–DEC cross-matching between the eCALIFA and Galaxy Zoo 2 datasets. As a result, the multimodal models were trained on a relatively small number of paired samples, which restricted their ability to learn more generalized representations. In addition, differences between the training and testing datasets may have affected the final classification accuracy. In addition, differences between the training and testing datasets may have affected the final classification accuracy, and the eight-class morphology mapping used to align the two surveys may have introduced some class-level ambiguity. Future work can focus on increasing the number of matched galaxy samples through larger spectroscopic surveys, improving dataset alignment procedures, and exploring alternative morphology mapping strategies. A larger and more balanced multimodal dataset would enable the models to learn richer spectral–morphological relationships and further improve classification performance and cross-dataset generalization.

Bibliography

[1] G. Martínez-Solaeche, R. García-Benito, R. M. González Delgado, Luis Díaz-García, S. F. Sánchez, A. M. Conrado, and J. E. Rodríguez-Martín, "Exploring Galaxy Properties of eCALIFA with Contrastive Learning."

[2] K. W. Willett et al., "Galaxy Zoo 2: Detailed Morphological Classifications for 304,122 Galaxies from the Sloan Digital Sky Survey."

[3] Jie Cao, Tingting Xu, Yuhe Deng, Linhua Deng, Mingcun Yang, Zhijing Liu, and Weihong Zhou, "Galaxy Morphology Classification Based on Convolutional Vision Transformer (CvT)."

[4] Guangping Li, Tingting Xu, Liping Li, Xianjun Gao, Zhijing Liu, Jie Cao, Mingcun Yang, and Weihong Zhou, "Galaxy Morphology Classification Using Multiscale Convolution Capsule Network."

[5] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang, "CvT: Introducing Convolutions to Vision Transformers."

[6] Chengming Hu, Xuan Li, Dan Liu, Haolun Wu, Xi Chen, Ju Wang, and Xue Liu, "Teacher-Student Architecture for Knowledge Distillation: A Survey."

[7] Jonathon Shlens, "A Tutorial on Principal Component Analysis."

[8] GeeksforGeeks, "Random Forest Algorithm in Machine Learning." Available online: https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/

[9] Soham Chattopadhyay, Arijit Dey, Pawan Kumar Singh, and Ram Sarkar, "DRDA-Net: Dense Residual Dual-Shuffle Attention Network for Breast Cancer Classification Using Histopathological Images."

[10] Wujie Zhou, Yuzhen Chen, Chang Liu, and Lu Yu, "GFNet: Gate Fusion Network With Res2Net for Detecting Salient Objects in RGB-D Images."

[11] Kitware,"Histogram-Based Thresholding." Available online: https://www.kitware.com/histogram-based-thresholding/

[12] Ankita Sarkar, Sarbani Palit, and Ujjwal Bhattacharya, "Demystifying Galaxy Classification: An Elegant and Powerful Hybrid Approach."

[13] Bartosz Bieganowski and Robert Slepaczuk ,"Supervised Autoencoder MLP for Financial Time Series Forecasting."

[14] S. F. Sánchez, L. Galbany, C. J. Walcher, R. García-Benito, and J. K. Barrera-Ballesteros, "The Calar Alto Legacy Integral Field Area Survey: Extended and Remastered Data Release," CALIFA Survey Data Release. Available: https://ifs.astroscu.unam.mx/CALIFA_WEB/public_html/CALIFA_ext_DATA_RELEASE.html Accessed: June 2026.

[15] S. F. Sánchez, L. Galbany, C. J. Walcher, R. García-Benito, and J. K. Barrera-Ballesteros, "The Calar Alto Legacy Integral Field Area (CALIFA)

[16] Astronn Documentation, "Galaxy10 DECaLS Dataset Documentation," Available: https://astronn.readthedocs.io/en/latest/galaxy10.html