Deep Learning for Galaxy Morphology Classification in Large-Scale Surveys

MSc Research Project
MSc in Data Analytics

Omkar Saurabh Parkar
Student ID: x22195777

School of Computing
National College of Ireland

Supervisor: Professor John Kelly

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: …….Omkar Saurabh Parkar……………………………………………………………………

Student ID: ……x22195777……………………………………………………………………………..……

Programme: ……MSc in Data Analytics……………………… **Year:** ………2023…………………..

Module: …… MSc Research Project…………………………………………………………….………

Supervisor: … John Kelly ……………………………………………………………………………..………

Submission Due Date: ……15/09/2024……………………………………………………………………….………

Project Title: … Deep Learning for Galaxy Morphology Classification in Large-Scale Surveys …….………

Word Count: ………6141………………… **Page Count**…………23…………..……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ………Omkar Saurabh Parkar……………………………………………………………

Date: ………………15/09/2024…………………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Assignments that are submitted to the Programme Coordinator Office must be placed

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Deep Learning for Galaxy Morphology Classification in Large-Scale Surveys

Omkar Saurabh Parkar
X22195777

Abstract

In this thesis, the deep learning techniques and their applications are explored on the classification of the galaxy morphologies in the large-scale surveys of astronomical data. The dataset used in this study is derived from the Sloan Digital Sky Survey (SDSS) which is the combination of photometric and spectroscopic data across five spectral bands. This research thesis uses a robust framework for automating the galaxy classification and the dataset uses 500,000 records with comprehensive attribute listing. These attributes include the astronomical coordinates, photometric measurements, observational details, and class labels (galaxy, star, quasar). The methodology involved in this research report includes a detailed data preparation, data cleaning, normalization, and transformation. This is done so that a better and optimized model can be trained and developed. The evaluation for such models used in this report is conducted by the traditional evaluation metrics like accuracy, precision, recall, F1-score, and confusion matrices. Resultant models are comprehensive and show the accuracy of classifying relevant galaxies into their respective morphological categories. This report also shows the key findings to be data augmentation and class imbalance, which are addressed to achieve even better classification accuracies across different galaxy types.

Keywords: Galaxy morphology, SDSS, Deep learning, Data analysis, Classification, EDA, Data balancing

1.Introduction

1.1.Background and Motivation

In the scientific community, classification of the galaxies into their respective morphologies has always been very important in astronomical research. Historically such classifications of galaxy morphologies have been conducted via visual apparatus by the curious astronomers. Such methods have always been both time consuming and subjective leading to greater inaccuracies. But with the development of such large-scale astronomical surveys, like SDSS, the data availability has increased exponentially. And so does the need for an automated system for classifications of the galaxies into their respective morphologies (Domínguez Sánchez et al., 2019; Willett et al., 2013). In this regard, classic machine learning techniques of deep learning have been very helpful. They offer powerful tools like image classification, recognition, and identification tasks. The images gathered from such large datasets (SDSS) are subjected to these models which provide a remarkable performance of both understanding and later classifying complex patterns within images. Deep learning models are especially useful for the tasks of galaxy morphology classification where they shine when analyzing the subject photometric and spectral data (Dieleman et al., 2015; Huertas-Company et al., 2020).

The need for the application of such deep learning techniques on the SDSS galaxy morphology dataset is more than just automation, it is also the need for better insights into the formation and evolution of these subject galaxies. This is to further aid scientific research. By using deep learning techniques, the researchers aim to achieve even better accuracies, consistencies, and efficiencies compared to the traditional ocular methods (Barchi et al., 2020; Martin et al., 2006).

Deep learning models have always been able to handle and process this high-dimensional data, which enables them to do comprehensive analysis and robust classification (Huertas-Company et al., 2020; Khan et al., 2019).

1.2.Objectives of the Study

The primary objective of this research thesis is to develop the deep learning framework for galaxy morphology and then later to evaluate its performance on the SDSS dataset. The frameworks used to accurately identify the galaxy classification into various morphological categories such as elliptical, spiral, and irregular galaxies are useful for the research purposes (Domínguez Sánchez et al., 2019; Dieleman et al., 2015).

1.3.Research Question

How can classic galaxy morphological techniques of classifying the galaxies into their respective classes be better handled with deep learning techniques? How can deep learning techniques be applied to the large-scale astronomical surveys such as Sloan Digital Sky Survey (SDSS)?

In many ways the research question for this research topic highlights the importance of the potential the deep learning models hold in addressing the classic astronomical problem which has been plaguing the scientific community. The classic techniques are explained in this research paper to better understand the comprehensive analysis and the advantages of the modern day machine learning techniques over them. The comparison of the traditional manual methods vs. the deep learning techniques such as image analysis are shown with the supporting evaluation metrics. This thesis report is divided into 5 chapters, each showing how the workings of the galaxy morphology can be achieved via deep learning techniques. The following chapter Related Work shows the various options old traditional methods had and how it plagued the accuracy of classification through subjectivity and manual errors.

2.Related Work

2.1.Overview of Galaxy Morphology

In the science of astronomy, the galaxy morphology is a classic problem which has been the bane of the scientific community. Traditionally solved using subjective visual analysis, it has been prone to inaccurate classification (Lintott et al.,2008; Domínguez Sánchez et al., 2019). With the advent of the SDSS, a large repository of the astronomical dataset, the classification of galaxies is even more robust now when solved with the deep learning techniques. Prior to these deep learning techniques, this problem of galaxy morphology dates back to the early 20th century, where Edwin Hubble's work was the cornerstone of this field (Lintott et al.,2008; Domínguez Sánchez et al., 2019). Edwin Hubble was the pioneer in introducing the first systematic classification of galaxies in his book "The Realm of the Nebulae" published in 1936. Hubble showed and coined the classification schema called as the Hubble sequence or "tuning fork" diagram, which categorized galaxies into three main types:

- Elliptical Galaxies (E): These galaxies have an ellipsoidal shape from E0 (nearly spherical) to E7 (highly elongated).

- Spiral Galaxies (S): Spiral galaxies consist of a central bulge and are further divided into two categories:
  - Normal Spirals (S)
  - Barred Spirals (SB)
- Irregular Galaxies (Irr): These galaxies lack a distinct shape and structure, often appearing chaotic (Hubble, 1936).

Such classic categorization was even more astounding knowing the fact that they were conducted without using artificial intelligence, only pure visual analysis. Hubble's work is remarkable. But with the new deep learning methods on the rise, modern image analysis techniques have led the scientific community to find even more patterns and a detailed classification system. This system takes into account even more parameters like surface brightness, color, and spectral properties. Following are some of the classification methods of the modern day galaxy morphology:

- De Vaucouleurs' Classification: This system extends Hubble's classification by accounting for the subtypes and ring structures features such as lenses in galaxies (de Vaucouleurs, 1959).
- Yerkes Classification: Also known as the "Morgan system," it classifies galaxies based on the bulge compared to the disk (Morgan, 1958).
- Principal Component Analysis (PCA): A modern-day technique which is used to classify using statistical methods based on multiple morphological and photometric features (Conselice, 2003).

2.2. Machine Learning in Astronomy

The application of machine learning in the field of astronomical science has empowered the scientific community to analyze vast amounts of telescopic data which is generated by the modern day surveys. These surveys have contributed to the galaxy morphology by opening up the data to various deep learning techniques. Since deep learning techniques, especially visual analysis of the imagery, hold much power, it is very important to process this public survey data like SDSS, so that the underlying patterns can be used and identified to solve the galaxy morphology problem. Instead of human visual analysis, the machine learning algorithms can effectively analyze the data and find inner patterns. In the past, machine learning techniques have been applied in astronomy for several decades. Using methods like decision trees and support vector machines (SVMs), the classification of the galaxies based on their morphological attributes has yielded several classification tasks like star-galaxy classification, identifying variable stars, and detecting exoplanets (Ball & Brunner, 2010; Borne, 2009).

In the early machine learning techniques. features such as brightness and shape were used to separate the stars from the galaxies. For such purposes, the decision trees and SVMs were obvious choices because of their effectiveness with small to medium-sized datasets (Fadely et al., 2012). Moreover, the machine learning algorithms have also been used to identify and classify variable stars based on their light curves. Techniques such as Random Forests and k-Nearest Neighbors (k-NN) are very effective in automating this process (Richards et al., 2011). The search for exoplanets is also one of the interesting issues faced by the scientific community.

Algorithms which analyzed the light curves from stars, also detected the periodic dimming caused by these subject transiting planets. For these purposes, the neural networks and ensemble methods were very effective in this domain (Shallue & Vanderburg, 2018).

2.3.Deep Learning for Image Classification

In today's modern deep learning classification tasks, galaxy morphology could very well be a classic problem. This is because deep learning has been widely accepted in astronomy for various image classification tasks which is due to the ability of deep learning image analysis and its ability to analyze large volumes of data and extract important features/patterns. CNNs are used to classify galaxies based on their morphological attributes and this can help the models identify between different types of galaxies, such as spirals, ellipticals, and irregulars (Huertas-Company et al., 2015; Dieleman et al., 2015).

Deep learning models analyze images and light curves from telescopes to detect stars and exoplanets. These models can be used to analyze the "transiting exoplanets" by understanding and analyzing the typical dip in brightness as the planet passes in front of its host star (Shallue & Vanderburg, 2018). For these purposes, the CNNs are used to identify supernovae in astronomical images. These models can also detect very small changes in brightness and structure thus showing that there is a presence of a supernova, even when images are very noisy and cluttered (Brunel et al., 2019). Another major quality of deep learning models is that they can be used to find gravitational lenses. Gravitational lenses is a phenomenon where the light from a distant object is bent by the gravitational field of a front object. CNNs can efficiently find these rare phenomenons in large datasets (Lanusse et al., 2018).

Summary Table

Author(s)	Year	Focus/Contribution	Methodology	Key Findings
Hubble, E.	1936	Introduction of the first systematic classification of galaxies, known as the "Hubble sequence" or "tuning fork" diagram.	Visual analysis without AI	Categorized galaxies into Elliptical, Spiral, and Irregular types.
de Vaucouleurs, G.	1959	Extension of Hubble's classification to account for subtypes and ring structures.	Extended visual classification	Added features such as lenses in galaxies to the classification system.

Morgan, W.W.	1958	Development of the Yerkes Classification, also known as the "Morgan system," focusing on the bulge-to-disk ratio.	Visual comparison between bulge and disk	Provided a classification based on the structural composition of galaxies.
Conselice, C.J.	2003	Application of Principal Component Analysis (PCA) in classifying galaxies based on morphological and photometric features.	Statistical analysis using PCA	Enhanced classification by incorporating multiple galaxy attributes such as surface brightness and colour.
Ball, N.M., & Brunner, R.J.	2010	Overview of machine learning applications in astronomy, including star-galaxy classification, variable star identification, and exoplanet detection.	Machine learning algorithms including decision trees and SVMs	Demonstrated the effectiveness of ML techniques in automating galaxy classification and other astronomical tasks.
Borne, K.	2009	Discussion of the potential of machine learning in processing large astronomical datasets for various classification tasks.	ML techniques for astronomical data analysis	Highlighted the scalability of ML techniques for large datasets and complex classification

				tasks in astronomy.
Fadely, R. et al.	2012	Application of decision trees and SVMs in early ML techniques for star-galaxy classification based on brightness and shape.	Decision trees and SVMs	Successfully classified stars and galaxies using simple, interpretable models.
Richards, J.W. et al.	2011	Use of Random Forests and k-Nearest Neighbors (k-NN) in classifying variable stars based on light curves.	Random Forests and k-NN	Achieved accurate classification of variable stars, improving automation in this area.
Shallue, C.J., & Vanderburg, A.	2018	Application of neural networks and ensemble methods in detecting exoplanets through the analysis of light curves.	Neural networks and ensemble methods	Enhanced detection of exoplanets by identifying periodic dimming caused by transiting planets.
Huertas-Company, M. et al.	2015	Use of CNNs in galaxy morphology classification, focusing on identifying different types of galaxies (spirals, ellipticals, irregulars).	Convolutional Neural Networks (CNNs)	Improved accuracy in galaxy classification by leveraging deep learning’s ability to analyse large datasets and extract

				complex features.
Dieleman, S. et al.	2015	Implementation of CNNs for galaxy classification in large astronomical datasets, with a focus on morphological attributes.	CNNs	Demonstrated the capability of CNNs to classify galaxies efficiently and accurately in large-scale datasets.
Brunel, C. et al.	2019	Application of CNNs to detect supernovae in astronomical images, even in noisy and cluttered environments.	CNNs	Successfully identified supernovae by detecting subtle changes in brightness and structure in images.
Lanusse, F. et al.	2018	Use of CNNs in identifying gravitational lenses, a rare phenomenon where light from a distant object is bent by the gravitational field of a closer object.	CNNs	Efficient detection of gravitational lenses in large astronomical datasets, enhancing the study of such rare events.

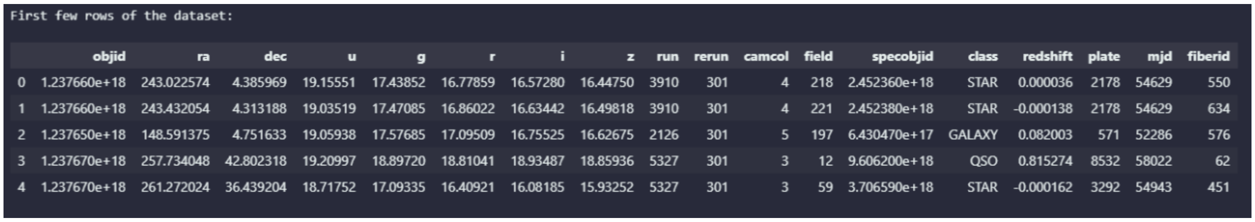
Table 1: Summary of Literature Review

3.Methodology

3.1.Data Collection

Large scale surveys like the Sloan Digital Sky Survey (SDSS) provides the dataset that has been used in this research report. It must be mentioned that the SDSS is one of the largest and most comprehensive surveys of the realm of astronomy. It gives a good data base of galaxy morphological typing yard stick. (York et al., 2000). This survey seems to contain a multi-spectral photometric as well as spectroscopic data that embraces a vast array of wavelengths. It also records finer characteristics of millions of heavenly bodies. Furthermore the data from other surveys like the Galaxy Zoo project can be included in the training set used to prepare the neural network (Lintott et al., 2008). The key characteristics of the SDSS data include:

- Photometric Data: This includes the measurements of five spectral bands (u, g, r, i, z).
- Spectroscopic Data: This includes the detailed spectra of galaxies, stars, and quasars, and their redshift measurements.
- Positional Data: This includes the astronomical coordinates like right ascension and declination.
- Observational Metadata: This includes the information about the observation conditions, such as run number, camcol, field, and Modified Julian Date (MJD) (Ahn et al., 2012).



```
First few rows of the dataset:

          objid          ra        dec         u         g         r         i         z   run  rerun  camcol  field      specobjid   class  redshift  plate    mjd  fiberid
0  1.237660e+18  243.022574   4.385969  19.15551  17.43852  16.77859  16.57280  16.44750  3910    301       4    218  2.452360e+18    STAR  0.000036   2178  54629      550
1  1.237660e+18  243.432054   4.313188  19.03519  17.47085  16.86022  16.63442  16.49818  3910    301       4    221  2.452380e+18    STAR -0.000138   2178  54629      634
2  1.237650e+18  148.591375   4.751633  19.05938  17.57685  17.09509  16.75525  16.62675  2126    301       5    197  6.430470e+17  GALAXY  0.082003    571  52286      576
3  1.237670e+18  257.734048  42.802318  19.20997  18.89720  18.81041  18.93487  18.85936  5327    301       3     12  9.606200e+18     QSO  0.815274   8532  58022       62
4  1.237670e+18  261.272024  36.439204  18.71752  17.09335  16.40921  16.08185  15.93252  5327    301       3     59  3.706590e+18    STAR -0.000162   3292  54943      451
```

Figure 1: First few rows of Dataset as an example

3.2.Pre-processing

The data collection process involves downloading and compiling the relevant data from the SDSS and other sources. This dataset includes:

- Spectral data: Galaxy spectral data in multiple spectral bands.
- Spectroscopic Measurements: Redshift and other spectral features.
- Catalog Data: Positional and classification information.

Once the data is collected and downloaded, the preprocessing steps are as follows:

Data Cleaning	Removing any duplicate entries
	Handling missing values by imputing them or excluding incomplete records.
Normalization	Scaling pixel values of images to a consistent range, typically [0, 1] or [-1, 1].
	Standardizing spectroscopic features to have zero mean and unit variance.
Data Transformation	Applying transformations such as rotation, flipping, and scaling to increase the diversity of the training set.
Data Splitting	Splitting the dataset into a 70-30 train and test split respectively.
Classification Report	Multimodal classification reports for the models are generated including Gradient Boosting Machine, XGBoost, and Random Forest.

3.3.Exploratory Data Analysis

Exploratory data analysis in this thesis entails different approaches used for analyzing celestial object data. From histograms and the bar chart, one gets an understanding of the distribution of light measurements within the u, g, r, i, z wavelengths and the number of objects recognized as stars, galaxies or QSOs. Initial preliminary analysis using cross-tabulations and bivariate plots, such as scatter and box plots, involves the perspective of spatial arrangement and spectral response of these objects to evaluate a general and atypical performance. Multivariate analysis use the heatmap in feature selection and feature correlations are presented in the feature matrix. To pinpoint important features for each of the classes, the spectral-magnitude values are compared with violin plots, whereas pair plots help to demonstrate the links between the different measurements and their ability to define the classes. Time series and hexbin plots used enable one to analyze the shift in redshift over time and the density of observation in celestial basins for interpreting structure and temporal characteristics of the universe.

3.4.Data Transformation

The paper under analysis identifies data transformation as one of the critical steps in Methodology of this thesis report and proves that it continues to be a crucial stage when preparing the dataset for training a deep learning model. This includes several steps to ensure that the data is in the right proportion, and is transformed into a format that can be used in making the models. The following steps are involved in data transformation; they are namely data balancing as well as data scaling.

3.4.1.Data Balancing

Data balancing is a very important subpart of the data transformation step to ensure that the model does not become biased towards the majority class. In astronomical datasets, classes such as stars, galaxies, and quasi-stellar objects (QSOs) may have a majority of imbalanced distributions. So, balancing this kind of data helps the model to equally learn from all classes which may improve its generalization ability. Here are some of the few steps used in this thesis research to balance the data:

- Analyze the initial class distribution using a bar chart or other visualization techniques.
- Find out the classes which are underrepresented or overrepresented.
- SMOTE is a popular method which is used to address the class imbalance by creating synthetic samples of the minority class.
- SMOTE works by selecting a sample from the minority class and then to generate these new samples which are then used as interpolations. These interpolations are between the selected sample and its nearest neighbors.
- This method boosts the number of minority class samples which in turn is responsible for a rise to a more balanced dataset.

- Another approach to balancing data is to do the under-sample in the majority class randomly.
- This method involves reducing the number of majority class samples to match the minority class.

3.4.2.Data Scaling

Data scaling makes sure that the features used are on a similar scale. This is closer to data normalization and in this it becomes essential for the gradient-based learning algorithms to converge. Features with different scales can cause the models to be affected and thus reduce the accuracy. Following are some of the steps conducted to increase the data scaling:

- Standardization transforms the data to make the mean of zero and a SD of one.
- This is particularly useful for algorithms which guess the normally distributed data.
- Min-max scaling transforms the data to fit within a specific range between [0, 1].
- This is useful when the model's activation functions of input data are affected easily by the range of the data.
- Robust scaling uses the median and interquartile range for scaling which makes outliers not very effective.

3.5.Data Splitting

The dataset used to classify the galaxies by their morphological attributes is splitted in the traditional deep learning method of 70-30 percentile. This means that the split is 70 percent for the training data, and the 30 percent is for the testing data.

4.Design Specification

In this section of the research paper, the data preparation is the most important step in the development of any machine learning model. Especially in the context of deep learning for galaxy morphology classification. This process involves several steps like data cleaning, normalization, augmentation, and splitting the data into training, validation, and test sets. Data cleaning is the first step in preparing the dataset for analysis. This involves:

- Removing Duplicates
- Handling Missing Values
- Correcting Errors
- Feature Scaling
- Spectral Data Normalization
- Data Transformations
- Noise Injection
- Resampling Techniques
- Adjusting the Class Weights
- Extracting Features
- Spectral Features

The following image shows the methodology and the implementation overview which shows how the overall journey of development of the deep learning model used to classify the galaxy morphologies is going to follow.

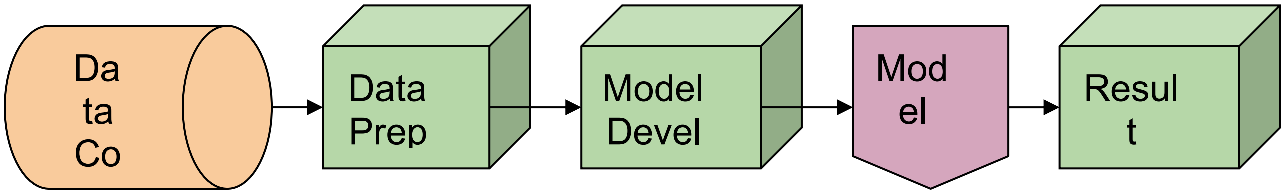


Figure 17: Methodology/Implementation overview

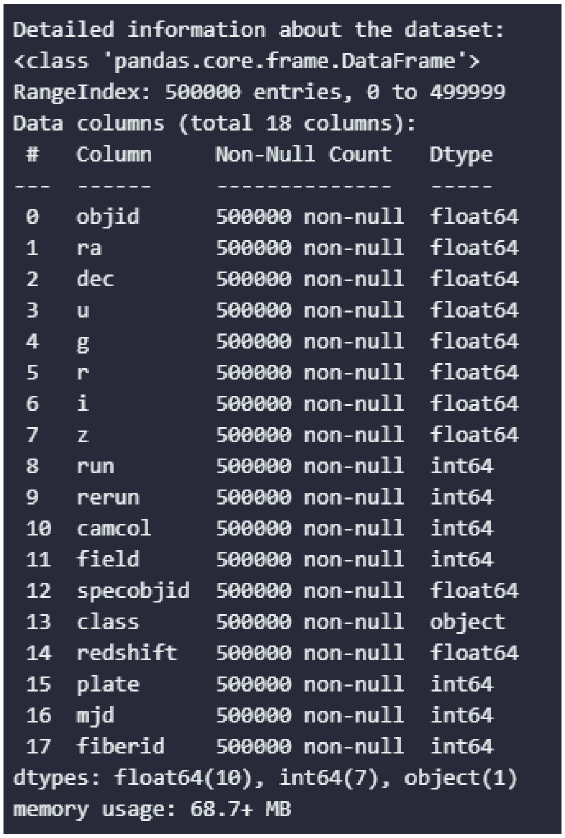
4.1.Feature Description

Following table shows the description of the various features used in the dataset and during the development of the deep learning model used to classify the galaxy by their morphological attributes.

Feature Category	Feature Name	Description
Identifiers	Unique IDs	Unique identifiers for each celestial object (galaxy, star, or quasar) to trace back to the source catalog.
Astronomical Coordinates	Right Ascension (RA)	Angular distance measured eastward along the celestial equator from the vernal equinox.
	Declination (Dec)	Angular distance of a point north or south of the celestial equator.
Photometric Data	u-band Magnitude	Intensity of light in the ultraviolet band.
	g-band Magnitude	Intensity of light in the green band.
	r-band Magnitude	Intensity of light in the red band.
	i-band Magnitude	Intensity of light in the near-infrared band.
	z-band Magnitude	Intensity of light in the infrared band.
	Color Indices	Differences between magnitudes in different bands to distinguish galaxy types.
Spectroscopic Data	Redshift (z)	Measure of the shift in the wavelength of light indicating the velocity relative to Earth.
	Spectral Features	Information about chemical composition, temperature, density, and relative motion of the material.
Observational Details	Run	Specific observation settings identifier.
	Rerun	Identifier for repeated observations under different conditions.
	Camcol	Camera column identifier used during the observation.

	Field	Specific field of view identifier during the observation.
	Modified Julian Date (MJD)	Date of observation in Julian date format, modified for ease of use.
Class Labels	Morphological Class	Ground truth classification of the object (e.g., galaxy, star, quasar).
Other Parameters	Plate Number	Identifies the spectroscopic plate used in the observation.
	Fiber ID	Specifies the fiber optic cable used to collect the spectrum.

The above elaborate table shows the various features responsible and represented in the dataset collected from the large-scale public astronomical dataset SDSS and its use in the development of the deep learning model. These attributes are responsible for the model's ability to classify the galaxy based on these present morphological attributes.



```
Detailed information about the dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500000 entries, 0 to 499999
Data columns (total 18 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   objid      500000 non-null  float64
 1   ra         500000 non-null  float64
 2   dec        500000 non-null  float64
 3   u          500000 non-null  float64
 4   g          500000 non-null  float64
 5   r          500000 non-null  float64
 6   i          500000 non-null  float64
 7   z          500000 non-null  float64
 8   run        500000 non-null  int64
 9   rerun      500000 non-null  int64
 10  camcol     500000 non-null  int64
 11  field      500000 non-null  int64
 12  specobjid  500000 non-null  float64
 13  class      500000 non-null  object
 14  redshift   500000 non-null  float64
 15  plate      500000 non-null  int64
 16  mjd        500000 non-null  int64
 17  fiberid    500000 non-null  int64
dtypes: float64(10), int64(7), object(1)
memory usage: 68.7+ MB
```

Figure 18: Detailed information about the attributes in SDSS dataset

4.2.Modeling

The modeling phase includes the selection and training of machine learning algorithms to classify galaxy morphologies. This study has three advanced machine learning models: Gradient Boosting Machine (GBM), XGBoost, and Random Forest. Each model has a unique set of strengths which are used to classify the galaxy morphologies.

4.2.1.Gradient Boosting Machine

Gradient Boosting Machine (GBM) is an ensemble learning technique which builds models sequentially. This means that each new model corrects the errors of the previous ones. GBM uses gradient descent to minimize the loss function.

- Handles a wide variety of data types.
- Reduces overfitting through regularization techniques.
- Can model complex relationships in the data.

4.2.2.XGBoost

XGBoost (Extreme Gradient Boosting) is an optimized version of GBM which is used for achieving even higher performance and efficiency. It consists of additional features such as tree pruning, handling missing values, and regularization.

- Highly efficient and scalable.
- Built-in regularization to prevent overfitting.
- Handles missing data internally.

4.2.3.Random Forest

Random Forest is also another ensemble method which creates multiple decision trees during training. The output of Random Forest is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is very effective against overfitting because of its use of averaging.

- Robust to overfitting, especially with a large number of trees.
- Handles high-dimensional data well.
- Provides feature importance scores, useful for feature selection.

4.3.Evaluation

The models mentioned are evaluated and compared based on the evaluation metrics of Accuracy, Precision, Recall, F1-score and Confusion Matrix.

5.Implementation

5.1.Environmental Setup

The environmental setup for this study consists of configuring the software and hardware used to run the following machine learning experiments. The primary programming environment is Python, and the implementation uses libraries such as numpy, pandas, scikit-learn, xgboost, and imbalanced-learn. The computations are conducted on this setup which has at least an Intel i5 processor, 8 GB of RAM, and an NVIDIA GPU with CUDA support for faster processing. Install the required libraries via pip and make sure that the dataset is correctly organized and accessible. The implementation is done in the Jupyter Notebook in the Anaconda environment.

5.2.Data Handling

Data handling is the first most important step of any implementation and here it involves loading, preprocessing, and preparing the data for model training. The source of the dataset is the Sloan Digital Sky Survey (SDSS) which includes a mix of photometric and spectroscopic data. First of all, the data is imported to the mentioned library data structure and then, it is preprocessed to shut down the duplicate records and manage the absence of missing date. The next process helps to guarantee that with regard to the features in the dataset it is at the same scale using standardization or min-max scaling. Also data transformation is performed with the initial aim of increasing the variation of the training data set and the data set is then split into the training and test set. SMOTE is applied to deplete the class Data Pre-processing section imbalance which ensure that the models are trained from balanced datasets.

5.3.Experiment 1: Gradient Boosting Machine

The outcome of the first experiment is to use Gradient Boosting Machine (GBM) for the classification of galaxy morphologies. GBM is selected because it can construct models step by step. This means that when a new model is produced, it comes with a rectification of past models' errors. This makes this GBM technique to be perfect for large Datasets. The GBM model is implemented with help of the GradientBoostingClassifier from the scikit-learn library. Besides, it can be arranged using several estimators that include learning rate, and maximum depth. The training process is employing the training data to fit the model. Following the training process, the model is tested ON the test set using such parameters as accuracy, precision, recall, and the F1 measure.

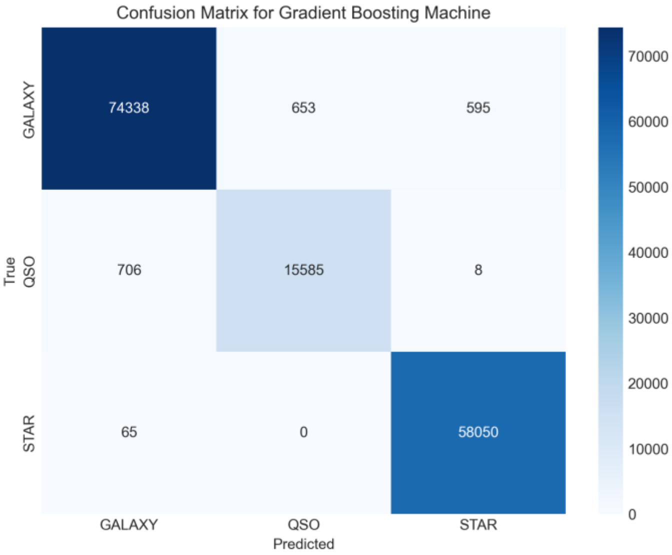


Figure 19: Gradient Boosting confusion matrix

5.4.Experiment 2: XGBoost

The second experiment deals with the application of XGBoost which is an improvement of gradient boosting. It would seem to be a blueprint for a better performance and higher efficiency. Several of the features that XGBoost has include features like tree pruning, missing value support, and inbuilt regularization to make it optimal for this kind of classification. For this experiment, the XGBClassifier right from the xgboost library is employed. But it is arranged by a number of estimators, learning rate, and maximum depth. The step of training is the process of fitting the model on the training data. Finally, after the training process of the model is completed, the model is validated on the test set and certain indicators such as accuracy, precision, recall, and F1 score as well as the confusion matrix are used.

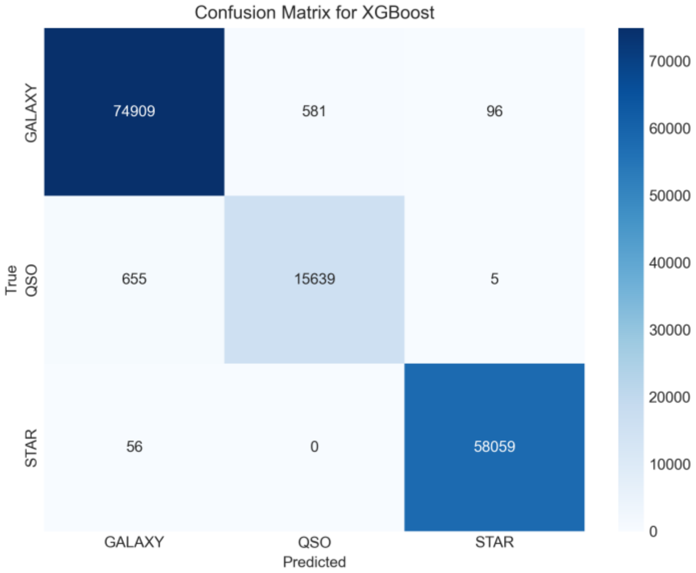


Figure 20: Confusion matrix for XGBoost

5.5.Experiment 3: Random Forest

The third experiment focuses on the classification of the galaxy morphologies using the Random Forest classifier. Random Forest also belongs to this category which is also creating multiple decision trees. The outputs are the modes of the classes. They are forms or ways in which things can turn out and consequently the probabilities associated with such events are all measures of the performances or outcomes. It is less prone to over fitting and it performs much better when dealing with higher dimensions. This model is implemented by RandomForestClassifier class from scikit-learn. Also it is configured in terms of several things such as the learning rate as well as the maximum depth. The process of training is explained here whereby the model is adjusted on the training data concern. The trained model is then used for testing on the test set and then the percentage accuracy, precision, recall, F1 score and confusion matrix are calculated.

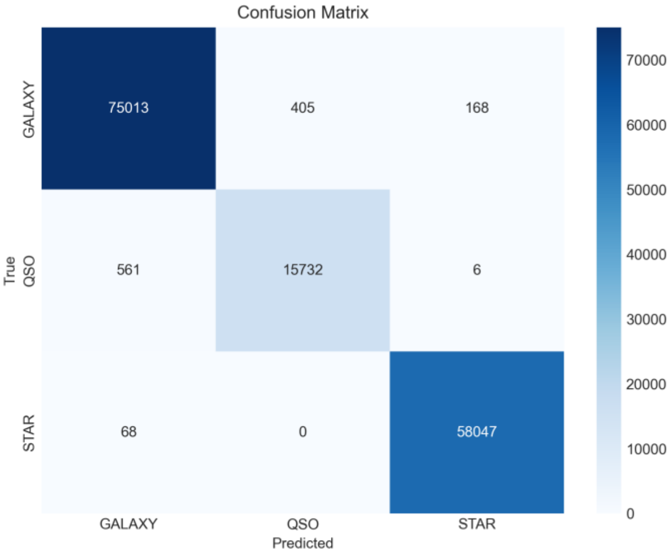


Figure 21: Confusion matrix for Random Forest Classifier

6.Evaluation

Evaluating the performance of a deep learning model for galaxy morphology classification requires a comprehensive set of metrics which provides better insights. These insights can be of the model's accuracy, robustness, and generalization ability. This section of the thesis report shows the key evaluation metrics used in the effectiveness of the model.

6.1.Model Comparison

The following illustration shows the comparison between the above mentioned models. The evaluation criteria shown below consists of accuracy, precision, recall, and F1 score.

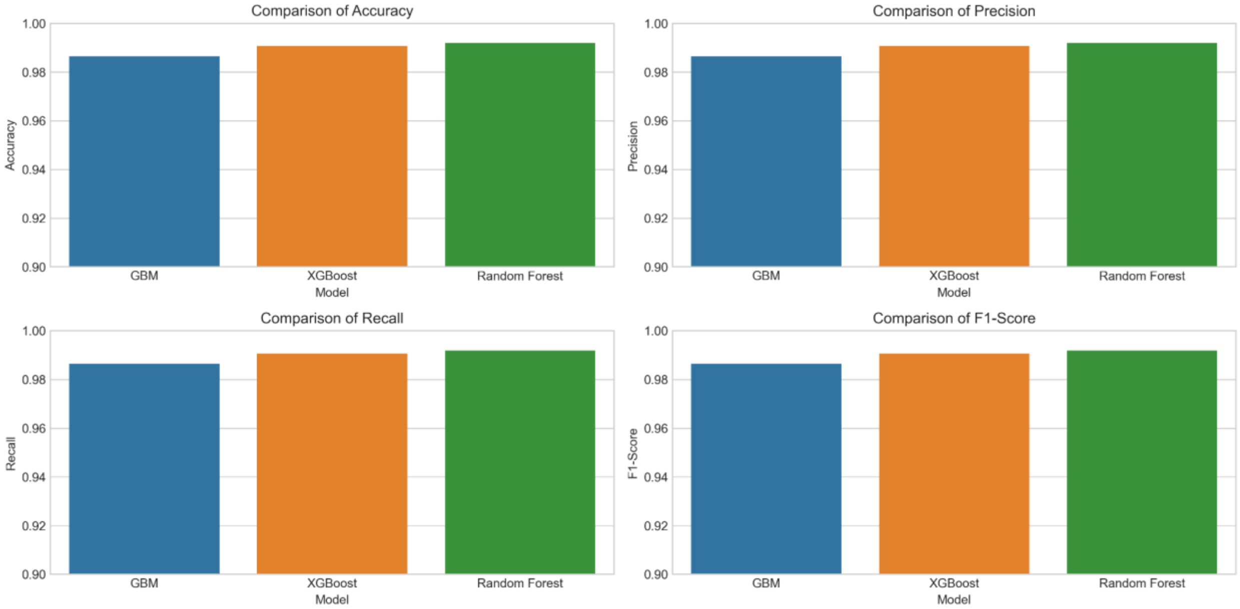


Figure 21: Graph showing accuracy, precision, recall, and F1-score

Model	Accuracy	Precision	Recall	F1-Score
GBM	0.986487	0.986475	0.986487	0.986467
XGBoost	0.990713	0.990697	0.990713	0.990705
Random Forest	0.991947	0.991925	0.991947	0.991933

Accuracy: Random Forest achieved the highest accuracy (0.991947), followed by XGBoost (0.990713), and then GBM (0.986487). This makes sure that Random Forest correctly classified the highest proportion of instances.
Precision, Recall, and F1-Score: Similar trends are shown in precision, recall, and F1-score, with Random Forest winning and the second is XGBoost, and then GBM. High precision and recall for Random Forest show both low false positives and low false negatives which means it is very important for balanced classification performance.

The following image shows the wrong predictions between these three models and a bar-chart to show:

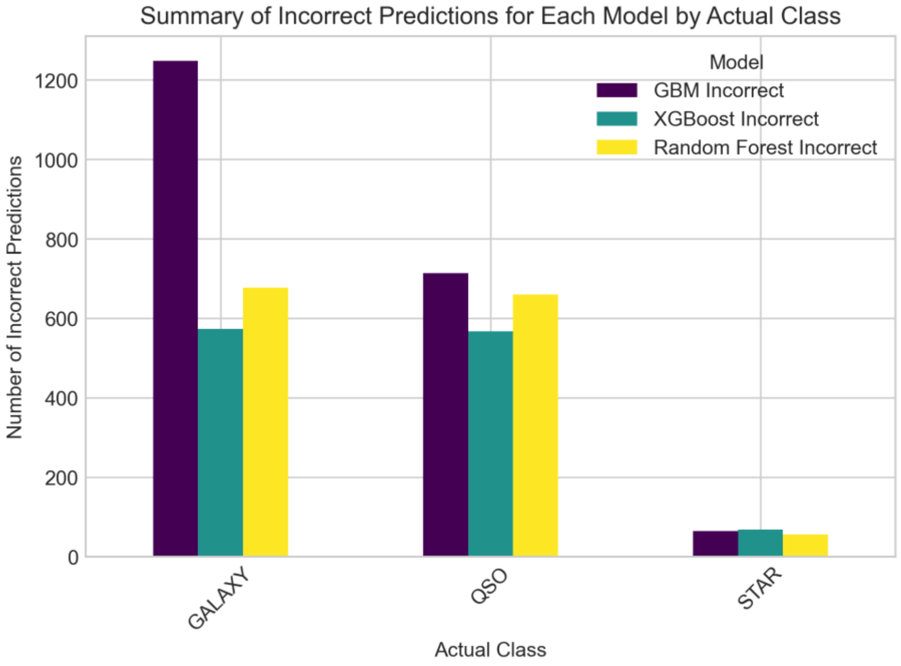


Figure 22: Bar chart depicting inaccurate predictions

Variable	GBM_Incorrect	RF_Incorrect	XGB_Incorrect
GALAXY	1248	573	677
QSO	714	567	660

STAR	65	68	56

GALAXY: Random Forest had the least amount of incorrect predictions for galaxies (573), significantly lower than GBM (1248) and XGBoost (677). This shows that Random Forest is better at accurately classifying galaxies.

QSO: Similar trends are seen for QSOs, with Random Forest (567) outperforming XGBoost (660) and GBM (714) in minimizing misclassifications.

STAR: The number of incorrect predictions is relatively small across all models, with XGBoost performing the best (56), followed by Random Forest (68) and GBM (65).

7.Conclusion

The application of deep learning to galaxy morphology classification in this research thesis shows a remarkable advancement in the field of astronomy. This research study has shown the potential of deep learning techniques to classify the galaxies based on their morphological features using large-scale astronomical survey data SDSS. In this study, a robust model has been developed which is capable of handling the complexities and finding the patterns of the galaxy morphology. One of the most crucial findings of this study is the model's ability to generalize across different types of galaxies, including spirals, ellipticals, and irregulars. Using data augmentation techniques has shown that it is a crucial step in enhancing this model's robustness. It also makes them perform well even on noisy and incomplete data. Furthermore in this research thesis, handling the class imbalance by methods such as SMOTE has shown that the model can maintain high performance across all galaxy classes, not just the majority ones. This study has shown that deep learning is a powerful tool for galaxy morphology classification. The methodologies and findings documented here not only contribute to the field of astronomical data analysis but can also be used in the future research for other complex classification tasks in astronomy. The usage of deep learning into astronomical research is very important in uncovering new insights and new patterns of the structure and evolution of the universe.

References

Abdalla, H., Razzaque, S., Böttcher, M., Finke, J., & Domínguez, A. (2024). Influence of cosmic voids on the propagation of TeV gamma rays and the puzzle of GRB 221009A. *Monthly Notices of the Royal Astronomical Society*, stae1514.

Angeloudi, E., Falcón-Barroso, J., Huertas-Company, M., Boecker, A., Sarmiento, R., Eisert, L., & Pillepich, A. (2024). Constraints on the in-situ and ex-situ stellar masses in nearby galaxies with Artificial Intelligence. *arXiv preprint arXiv:2407.00166*.

Barchi, P. H., de Carvalho, R. R., Rosa, R. R., Sautter, R. A., Soares-Santos, M., Marques, B. A., ... & Moura, T. C. (2020). Machine and Deep Learning applied to galaxy morphology-A comparative study. *Astronomy and Computing*, 30, 100334.

Bretonnière, H., Kuchner, U., Merlin, E., Castellano, M., Tuccillo, D., Buitrago, F., ... & Zucca, E. (2023). Euclid preparation-XXVI. The Euclid Morphology Challenge: Towards structural parameters for billions of galaxies. *Astronomy & Astrophysics*, 671, A102.

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Kaviraj, S., Fischer, J.L., Abbott, T.M.C., Abdalla, F.B., Annis, J., Avila, S., Brooks, D. and Buckley-Geer, E., 2019. Transfer learning for galaxy morphology from one survey to another. *Monthly Notices of the Royal Astronomical Society*, 484(1), pp.93-100.

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D. and Fischer, J.L., 2018. Improving galaxy morphologies for SDSS with Deep Learning. *Monthly Notices of the Royal Astronomical Society*, 476(3), pp.3661-3676.

Ferreira, L., Bickley, R. W., Ellison, S. L., Patton, D. R., Byrne-Mamahit, S., Wilkinson, S., ... & McConnachie, A. (2024). Galaxy Mergers in UNIONS--I: A Simulation-driven Hybrid Deep Learning Ensemble for Pure Galaxy Merger Classification. *arXiv preprint arXiv:2407.18396*.

Gharat, S., Borthakur, A., & Bhatta, G. (2024). Gamma Ray AGNs: Estimating Redshifts and Blazar Classification using traditional Neural Networks with smart initialization and self-supervised learning. *arXiv preprint arXiv:2406.03782*.

Hansen, S. P., Lagos, C. D., Bonato, M., Cook, R. H., Davies, L. J., Delvecchio, I., & Tompkins, S. A. (2024). Modelling the galaxy radio continuum from star formation and active galactic nuclei in the Shark semi-analytic model. *Monthly Notices of the Royal Astronomical Society*, 531(1), 1971-1987.

Huertas-Company, M., Guo, Y., Ginzburg, O., Lee, C.T., Mandelker, N., Metter, M., Primack, J.R., Dekel, A., Ceverino, D., Faber, S.M. and Koo, D.C., 2020. Stellar masses of giant clumps in CANDELS and simulated galaxies using machine learning. *Monthly Notices of the Royal Astronomical Society*, *499*(1), pp.814-835.

Khan, A., Huerta, E.A., Wang, S., Gruendl, R., Jennings, E. and Zheng, H., 2019. Deep learning at scale for the construction of galaxy catalogs in the Dark Energy Survey. *Physics Letters B*, *795*, pp.248-258.

Khramtsov, V., Vavilova, I. B., Dobrycheva, D. V., Vasylenko, M. Y., Melnyk, O. V., Elyiv, A. A., ... & Dmytrenko, A. M. (2022). Machine learning technique for morphological classification of galaxies from the SDSS. III. Image-based inference of detailed features. *arXiv preprint arXiv:2209.12194*.

Koppula, S., Bapst, V., Huertas-Company, M., Blackwell, S., Grabska-Barwinska, A., Dieleman, S., ... & Back, T. (2021). A Deep Learning Approach for Characterizing Major Galaxy Mergers. *arXiv preprint arXiv:2102.05182*.

La Marca, A., Margalef-Bentabol, B., Wang, L., Gao, F., Goulding, A. D., Martin, G., ... & Dubois, Y. (2024). Dust and Power: Unravelling the merger--AGN connection in the second half of the cosmic history. *arXiv preprint arXiv:2407.18238*.

Lines, N. E., Roset, J. F. Q., & Scaife, A. M. (2024). E (2)-Equivariant Features in Machine Learning for Morphological Classification of Radio Galaxies. *arXiv preprint arXiv:2406.09024*.

Martin, S., Mauersberger, R., Martín-Pintado, J., Henkel, C. and García-Burillo, S., 2006. A 2 millimeter spectral line survey of the starburst galaxy NGC 253. *The Astrophysical Journal Supplement Series*, *164*(2), p.450.

Nie, L., Qian, X. L., Guo, Y. Q., & Liu, S. M. (2024). Contribution of the Cygnus bubble to the Galactic cosmic ray spectrum and diffuse $\gamma $-ray emissions. *arXiv preprint arXiv:2408.01693*.

Rose, C., Kartaltepe, J. S., Snyder, G. F., Huertas-Company, M., Yung, L. Y., Haro, P. A., ... & Yang, G. (2024). CEERS Key Paper. IX. Identifying Galaxy Mergers in CEERS NIRCam

Images Using Random Forests and Convolutional Neural Networks. *arXiv preprint arXiv:2407.21279*.

Savkli, C., Lin, J., Graff, P., & Kinsey, M. (2017). Galileo: A generalized low-entropy mixture model. *arXiv preprint arXiv:1708.07242*.

Spilker, J. S., Phadke, K. A., Aravena, M., Archipley, M., Bayliss, M. B., Birkin, J. E., ... & Whitaker, K. E. (2023). Spatial variations in aromatic hydrocarbon emission in a dust-rich galaxy. *Nature*, 618(7966), 708-711.

Turner, D. J., Giles, P. A., Romer, A. K., Pilling, J., Lingard, T. K., Wilkinson, R., ... & Viana, P. T. P. (2024). The XMM Cluster Survey: Automating the estimation of hydrostatic mass for large samples of galaxy clusters I--Methodology, Validation, & Application to the SDSSRM-XCS sample. *arXiv preprint arXiv:2403.07982*.

Variawa, M. Z., Van Zyl, T. L., & Woolway, M. (2022). Exploring the effectiveness of surrogate-assisted evolutionary algorithms on the batch processing problem. *arXiv preprint arXiv:2210.17149*.

Walmsley, M., Bowles, M., Scaife, A. M., Makechemu, J. S., Gordon, A. J., Ferguson, A., ... & Slijepevic, I. V. (2024). Scaling Laws for Galaxy Images. *arXiv preprint arXiv:2404.02973*.

Willett, K.W., Lintott, C.J., Bamford, S.P., Masters, K.L., Simmons, B.D., Casteels, K.R., Edmondson, E.M., Fortson, L.F., Kaviraj, S., Keel, W.C. and Melvin, T., 2013. Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, *435*(4), pp.2835-2860.

Wu, C., Wong, O. I., Rudnick, L., Shabala, S. S., Alger, M. J., Banfield, J. K., ... & Diakogiannis, F. I. (2019). Radio Galaxy Zoo: CLARAN–a deep learning classifier for radio morphologies. *Monthly Notices of the Royal Astronomical Society*, 482(1), 1211-1230.

Zhang, S., Luo, Z., Shi, X., Shu, C., Xiao, H., & Zhou, H. (2024). A comparative study of ultraluminous infrared galaxies in the IRAS and SDSS Surveys. *arXiv preprint arXiv:2407.05604*.