Image Processing and Deep Learning for Space Telescope Images: An Exploratory Data Analysis Approach

by

Md. Mahir Jawad
20101285
Omar Ibtesam Nibir
20101293
Riana Islam Shawon
20101374
Syed Asif Mahbub
19101451

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
January 2024

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Mahir Jawad
20101285

Omar Ibtesam Nibir

Omar Ibtesam Nibir
20101293

Riana Islam Shawon

Riana Islam Shawon
20101374

Syed Asif Mahbub

Syed Asif Mahbub
19101451

Approval

The thesis/project titled "Image Processing and Deep Learning for Space Telescope Images: An Exploratory Data Analysis Approach" submitted by

1. Md. Mahir Jawad (20101285)
2. Omar Ibtesam Nibir (20101293)
3. Riana Islam Shawon (20101374)
4. Syed Asif Mahbub (19101451)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science.

Examining Committee:

Supervisor:
(Member)

Riad Ahmed
Lecturer
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

Exposing patterns in cosmic affairs from space telescope images is performed by a strong methodology configured by Exploratory Data Analysis (EDA) and advanced image processing techniques. This research utilizes machine learning to improve the categorization of stars and galaxies that space telescopes have taken images of, and it uses EDA to uncover hidden phenomena in celestial images. Automated categorization is essential for the purpose of improving relevant research due to the large quantity of astronomical datasets from missions such as the Sloan Digital Sky Survey (SDSS). Having applied contrast enhancement and noise reduction to the images, we use an optimized ResNet50 model for binary classification. In particular, The ResNet50 model which is pre-trained on the ImageNet database was altered to distinguish between stars and galaxies. The results demonstrate how potently image processing and deep learning work together, as seen by the remarkable classification accuracy of around 95%. This method illustrates how machine learning could potentially be used to automate astronomical data analysis, advancing astroinformatics and providing scalable solutions for the space missions to come.

Keywords

Exploratory Data Analysis (EDA); Image Processing; Machine Learning; Space Telescopes; Sloan Digital Sky Survey (SDSS); Contrast Enhancement; Noise Reduction; ResNet50; ImageNet; Binary Classification; Astroinformatics; Astronomical Data Analysis; Automated Categorization

Table of Contents

1 Introduction **1**
1.1 Background . . . 1
1.2 Problem Statement . . . 2
1.3 Research Objectives . . . 3

2 Literature Review **5**

3 Methodology **10**
3.1 Workflow Diagram . . . 11
3.2 Description of Data . . . 12
3.2.1 Collection and Preparation of Data . . . 12
3.2.2 Dataset Overview and Image Characteristics . . . 15
3.3 Data Preparation . . . 17
3.4 Data Pre-Processing . . . 23
3.4.1 Data Extraction and Organization . . . 23
3.4.2 Image Enhancement Using Contrast-Limited Adaptive Histogram Equalization (CLAHE) . . . 24
3.4.3 Image Resizing . . . 24
3.4.4 Image Normalization . . . 25
3.4.5 Data Augmentation . . . 25
3.4.6 Class Label Encoding . . . 26
3.4.7 Splitting the Dataset and Generating CSV Files . . . 27
3.4.8 Combining Dataset . . . 27

4 Analysis **28**
4.1 Performance Measure . . . 28
4.1.1 Evaluation Metrics . . . 28
4.1.2 Confusion Matrix . . . 29
4.1.3 Model Summary . . . 29
4.2 Comparison Analysis . . . 30
4.2.1 Accuracy Comparison . . . 30
4.2.2 Model Architecture Comparison . . . 31
4.2.3 Classification Report Comparison with traditional machine learning algorithms . . . 32
4.2.4 Enhancement of Model Performance through Integrated Feature Sets . . . 32
4.3 Discussion . . . 34

5 Conclusion **35**
5.1 Future Work 35

Bibliography **36**

List of Figures

3.1 Workflow Diagram 11
3.2 Sample Galaxy Images 12
3.3 Sample Star Images 12
3.4 Custom ResNet50 Model Summary 13
3.5 Class Distribution of Stars and Galaxies 18
3.6 Distributions of RA, DEC, and Redshift from the SDSS Star-Galaxy Dataset 19
3.7 Photometric Analysis of u, g, r, i, z Magnitudes: Scatter Plots and Histograms 20
3.8 Visualization of the Combined Feature Vector 27

4.1 Accuracy and Loss Curves of Custom ResNet50 28
4.2 Comparison of performance metrics across traditional ML models . . 29
4.3 Custom ResNet50 30
4.4 Custom CNN 30
4.5 Custom ResNet50 Architecture 31
4.6 CNN Architecture 31
4.7 Combined Features Model Accuracy and Loss Curves 32
4.8 Image Features Accuracy and Loss Curves 33

List of Tables

4.1 Classification Report 29
4.2 Comparison of Performance Metrics Across Different Models 32
4.3 Performance Comparison 33

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

AGN Active Galactic Nucleus

AHE Adaptive Histogram Equalization

CLAHE Contrast Limited Adaptive Histogram Equalization

CNN Convolutional Neural Network

CSV Comma-Separated Values

DEC Declination

DL Deep Learning

EDA Exploratory Data Analysis

GAP Global Average Pooling

ML Machine Learning

PCA Principal Component Analysis

RA Right Ascension

ReLU Rectified Linear Unit

$ResNet50$ Residual Network with 50 layers, a model used for deep learning

RGB Red, Green, Blue (color model)

SDSS Sloan Digital Sky Survey

Chapter 1

Introduction

The cosmos, an infinite tapestry of galaxies and celestial marvels, has endlessly fascinated human curiosity, propelling scientific exploration. The emergence of space telescopes heralds a transformative epoch, presenting awe-inspiring images of distant cosmic realms. Beyond their visual splendor, these images serve as intricate datasets teeming with astronomical information. This paper immerses itself in the realm of "Exploratory Data Analysis and Image Processing of Space Telescope Images," with the objective of unraveling concealed narratives through advanced data analysis. The challenge lies in distilling meaningful insights from complex and often noisy images, surpassing conventional astronomical analysis methods. Much like the progress in modern cancer research where advancements in immuno-therapy, precision medicine, health equity, liquid biopsies, and machine learning confront the intricacies of cancer, our exploration navigates the complexities of celestial imagery to deepen our comprehension of the universe.

1.1 Background

Our understanding of the universe is broadened through astronomical images which helps us in analyzing detail information regarding all celestial objects. Different space telescopes like James Webb Telescope, Hubble's Telescope and sky data collector platforms like Sloan Digital Sky Survey (SDSS) and GalaxyZoo offers major information about the past,present,future and evolution of our cosmos. But many obstacles are faced due to the collection of a large number of data: as the images does not possess much information to the naked eyes so researchers must develop accurate and efficient methods to classify and extract meaningful patterns and connections from these vast image datasets. It is not possible to accurately annotate these images manually as this is very time consuming and inefficient. To solve this problem, machine learning, or in specific deep learning became popular so that such type of classification works are automated and the report is accurate too. Convolutional neural networks (CNNs) is one of the main component of machine learning which helps in classifying visual features of an image in such ways that is too difficult for manual human labour. Though such innovative ways exists, still some challenges in image quality like low contrast,lighting and noise creates obstacles in achieving desired accuracy from ML models. To solve these challenges, this research applies advanced image preprocessing techniques, to enhance image quality by improving contrast and reducing noise. A fine-tuned deep learning model is then trained for binary classification of stars and galaxies. Additionally, data augmentation techniques are used to increase the diversity of the training data, enabling better feature extraction of unseen images. The result of this research is a proof that not only is the method of combining preprocessing, deep learning, and dataset modification efficient for astronomical classification automation, but it also provides a scalable solution for improving classification in future works. As a method to evolve astroinformatics, such an approach leads the way to more efficient large-scale analysis of space data by opening new avenues for exploration of the enormous collection of data from the universe.

1.2 Problem Statement

Classifying the stars and galaxies is among the initial steps to undertake in any attempt to start to understand the nature of the universe's structure-and, by extension, the origin of the heavenly bodies themselves. The very research problem the study attempts to address is how deep knowledge can be mined from the visually rich but complex snapshots of the heavens taken by the space telescopes. While such images are imposingly spectacular and complex, they form a formidable barrier for traditional methods of analysis. The overlying question that sets the basis for this research is: "How can an automated, scalable, and accurate method be developed to classify stars and galaxies from space telescope images and extract the relevant metadata from such images while overcoming challenges in image quality degradation and variability of the celestial objects being imaged?
Stars and galaxies, although similar in their general outlooks and classifications, differ in the range of shapes and sizes to brightness levels and many other factors. Galaxies may take the form of spirals, ellipticals, or irregulars; and depending on either their orientation or their distance from Earth, their appearance can become very different. Stars also vary in brightness and type, and sometimes overlap the images of galaxies, which makes their identification a little more complicated. Class imbalance is another issue which is common because the number of images of different celestial objects in the dataset are not always equal which leads models to be biased toward more frequent classes. Extracting meaningful features from these complex images also requires advanced deep learning algorithms. Traditional machine learning methods like Decision Tree Classifier, Naive Bayes and Logistic Regression often struggle to capture intricate diverse features, and models must be trained to recognize these details through effective data processing. The vast volume of data also presents logistical barriers, as the millions of images generated by astronomical research centres requires massive computational resources for storage, processing, and analysis. Deep learning models like convolutional neural networks (CNNs) has substantial power and memory, especially for high-resolution images which makes it capable of optimizing calculations for both efficiency and accuracy. While deep learning models like ResNet50 have proven successful in astronomical classification, concatenation of image metadata and visual features along with different preprocessing techniques are essential for improving input data quality. Without such enhancement, even the most dense and complex ML models may struggle to identify simple and key astronomical features. This study aims to enhance image quality using popular image preprocessing techniques, customize the trainable dataset and utilize ResNet50 to improve classification accuracy.

1.3 Research Objectives

The main goal of this research is to create an automated, efficient, and accurate method for classifying celestial objects like stars and galaxies from space telescope images. As large-scale astronomical surveys collect millions of high-resolution images, traditional manual classification methods are no longer practical. The research uses advanced image processing and machine learning techniques to tackle challenges in classifying these vast datasets. The key objectives are:

Enhance space telescope image quality with preprocessing techniques: Issues like noise, low contrast, and uneven lighting can obscure important features in space telescope images. This research applies contrast-limited adaptive histogram equalization (CLAHE) to improve image clarity, ensuring essential details are retained while reducing noise. These enhanced images form a reliable foundation for training machine learning models.

Implement a deep learning model for celestial object classification: This research will use convolutional neural networks (CNNs), specifically the ResNet50 model, to classify stars and galaxies. By fine-tuning the pre-trained ResNet50 architecture for binary classification, the study optimizes layers, parameters, and the training process to balance high accuracy with computational efficiency.

Address class imbalance in astronomical datasets: Astronomical datasets often have more stars than galaxies, leading to class imbalance. Techniques like class weighting and data augmentation will be used to ensure the model performs well on both classes, especially the minority class (galaxies).

Use data augmentation to improve model generalization: Data augmentation introduces variability into the training set by applying transformations like rotation, zooming, flipping, and brightness adjustments. These techniques help the model generalize to unseen data and avoid overfitting.

Model Merging to improve dataset features: The image dataset model is concatenated with metadata features of the celestial objects to create a more advanced and informative dataset which will help the model to connect and predict the outcome based on correlated features.

Evaluate model performance using multiple metrics: The model will be assessed using accuracy, precision, recall, F1-score, confusion matrices, and classification reports. These metrics will measure the model's ability to correctly classify stars and galaxies, especially with imbalanced datasets.

Optimize computational efficiency: With large datasets, efficient processing is crucial. This research streamlines the entire pipeline, from image preprocessing to model training, aiming for both high accuracy and reduced computational resource use.

Contribute to astroinformatics by developing scalable, automated solutions: The study aims to create scalable solutions for future astronomical missions, such as those of the James Webb Space Telescope. By automating celestial object classification, the research contributes to the broader goal of automating astronomical data analysis.

Explore future applications beyond binary classification: While focused on binary classification (stars vs. galaxies), this research lays the foundation for more complex tasks like multi-class classification and anomaly detection. It demonstrates the adaptability of the developed techniques to various astronomical classification challenges.

By achieving these objectives, the research aims to show that advanced preprocessing and deep learning models can significantly enhance celestial object classification, making it scalable and accurate for current and future large-scale astronomical surveys.

Chapter 2

Literature Review

Savyanavar et al.[1], in their work, focus on the classification of star and galaxy images using some ML and DL algorithms. They first try a number of traditional ML algorithms like RF, KNN, and SVM, and achieved an accuracy of 78.68% using RF as the most successful. They then present a new architecture for a CNN, which includes several convolution and fully connected layers. The features are extracted and refined from the ARIES dataset of 3986 images of stars and galaxies. Contrarily, a more accurate output using this CNN model is 92.44%, hence promising better handling of image data than other techniques. Architecture consists of introducing some data augmentation techniques: rotation, translation, and others to avoid overfitting. It is clear from these results that automatically extracted relevant features by CNN give better classification performance and make it a promising tool for most astronomical image analyses. Enhancements could be added with more datasets being added to the model or refining the model for real-time applications in cosmology.

In the work of Emma Leifer[2], the classification of such astronomical objects as stars and galaxies is discussed by using different machine learning models. The study assesses the performance of the following four models: a CNN, logistic regression, random forest classifier, and a small neural network, paying greater attention to their performance without additional measurements concerning brightness or size. It includes 3,986 images of size 64×64 pixels preprocessed by PCA. Among the models compared, the neural network achieved the best accuracy of 84% out of sample. Given the amount of data provided by modern telescopes, LSST will catalog approximately 40 billion images, this paper insists on the relevance of machine learning for efficient data processing. It also refers to previous related work in this field and states that with machine learning, the time to correctly classify the objects in space can be significantly reduced, thereby increasing our knowledge of those objects and their characteristics.

The research of Ganesh Ranganath Chandrasekar Iyer and Krishna Chaithanya Vastare [3] centered on developing a deep learning network to automate the classification of celestial objects. They tackled the problem of traditional methods' key limitation: the requirement of a large amount of time and energy from human experts to carry out the work of necessary classification. Their study's focus was on data that came from the famous Sloan Sky Survey and on examining the features of about 25,000 objects, around half of which were on the boundary of being classified as stars or galaxies. They developed a convolutional neural network (CNN) with eleven layers that can be trained. These layers use leaky ReLUs as activation functions and employ dropout for overfitting reduction. The model was evaluated with respect to the area under the ROC curve, achieving an impressive score of 0.97 on a test set of 1,500 instances. Compared with other existing classification methods, this performance validates the effectiveness of the architecture. Future work will entail scaling their model up to larger datasets and investigating its performance compared with other machine learning methods. Another issue they intend to solve is the limits of the hardware they used when training neural networks. They hope to highlight the prospects of convolutional neural networks for accurately classifying astrophysical objects with a minimum of human supervision.

In the study led by Khalifa et al.[4],a deep learning architecture relying on convolutional neural networks is proposed to categorize galaxies into three principal types: elliptical, spiral, and irregular. The complex model incorporates eight layers including an initial convolutional layer employing 96 filters to extract distinguishing features, followed by two fully connected layers performing the ultimate classification. When trained on a dataset of 1,356 galaxy photographs sourced from the EFIGI catalog, the model realized a highly impressive testing accuracy of 97.272%. The researchers underscore that traditional manual classification strategies are inefficient due to the immense volumes of astronomical data generated by modern telescopes. They compare their findings with previous work, demonstrating superior results exceeding existing methodologies that achieved accuracies ranging from only 79% to 93%. Automated galaxy classification proves pivotal in advancing our comprehension of celestial bodies, expediting exploration into how galaxies arise and change throughout time. Much remains to be done to refine the model, enhancing its strength when handling bigger archives of images and facts, ensuring astronomers can leverage its abilities to continue solving mysteries of the cosmos. The significant paper underscores applying the technique to more information, confirming findings across a wealth of astronomical observations.

In the paper, Urechiatu and Frincu [5] proposed a new architecture of CNN to classify morphologies of galaxies better by using a dataset comprising 10,000 images from the Galaxy Zoo 2 project. The classification of galaxies in this study used was into five classes: completely round smooth, in-between smooth, cigar-shaped smooth, edge-on, and spiral. Extensive pre-processing techniques were followed, involving enhancement and data augmentation techniques to avoid overfitting, which in total resulted in 120,000 training samples. The performance of the CNN was extensively tested against well-established architectures such as ResNet-50 and DenseNet. A remarkable accuracy of 96.83% was achieved, far outperforming all existing models. It also points out the problem of class imbalance in the dataset, especially for some classes, and allows the fact that future research can further fine-tune the methods of classification. This work will automate the classification of galaxies so as to reduce the manual work done by astronomers; hence, it will be a contribution to the field of astrophysics with more advanced data analysis techniques.

For the classification between AGN and star-forming galaxies, Brandon Xu [6] used a CNN; for this purpose, the NASA Sloan dataset has been utilized. Given the traditional spectroscopic methods that are really expensive and time-consuming, this research aims at amending these shortages by employing machine learning to help with automated classification. It is a CNN model trained on more than 105,000 images of successive convolution layers, extracting higher-level features of the input data. It yields an accuracy of about 78.1% on a test set of 14,130 images, with an impressively high AUC score of 0.85, indicating excellent performance in distinguishing between the two galaxy types. The study further points out the challenges related to data imbalance and suggests that further refinements in hyperparameters might lead to better classification accuracy. Thus, Xu's work proves that CNN can serve as a very feasible alternative in galaxy classification and hence facilitate newer dimensions in astronomy data analysis in future research undertakings.

Zhaoyan Xu et al. [7] proposed a custom RS CNN for the classification of spiral galaxies and stars, poised at the heart of inefficiency in methods traditionally used in astronomy. The paper points out an exponential growth of astronomical image data with the help of emerging digital facilities in astronomy and further says that conventional methods of machine learning cannot cope with such big datasets and, hence deep learning should be used instead. It optimizes the RS CNN model through mainstream architectures and reaches a great accuracy of 99.5%, with an AUC score of 0.996, which outperforms the SENet model by a big margin of about 9.7% in accuracy, with drastically reduced parameters and computation cost of the model. The work has highlighted the importance of various image preprocessing techniques in improving model performance and preventing overfitting, such as affine transformations and normalizations. In all, this review portrays the efficiency of deep learning algorithms in classifying astronomical images and consequently lays a foundation for further advances in galaxy morphology analysis using automation methods; hence, a proof that there can be higher efficiency in processing and analyzing astronomical data.

Nicholas M. Ball et al. [8] use a decision tree algorithm to classify approximately 143 million nonrepeat photometric objects from the Sloan Digital Sky Survey Third Data Release into star and galaxy candidates. They included a new class for categorization: "neither star nor galaxy," or nsng for short. This ensures that the high accuracy of the classifications at the fainter magnitudes is improved compared to traditional methods of star/galaxy separation. The decision tree technique, run on a training set of 477,068 objects with spectroscopic data, enables the robust classification of a very large data set: the classification of about 22 million objects for $r < 20$ should be of expected reliability. The authors explain their methods, which include refining the decision tree parameters with supercomputing resources, and are able to supply probabilistic classifications, allowing users to take their own choice of selection criteria. Blind tests on a number of surveys confirm that their method is efficient by showing that their classification is still robust even beyond the spectroscopic limits. The work is quite a step forward in the automation of astronomical classification, providing a publicly available data set that enables anyone with an interest in astronomy to do research and analysis. In summary, this study by Ball et al. underlines the capability of machine learning methods in their ability to handle large-scale astronomical data highly effectively and accurately.

Edward J. Kim and Robert J. Brunner [9] have proposed a deep convolutional neural network framework for star-galaxy classification that overcomes the shortcomings of traditional classifiers, which need features to be manually extracted from astronomical catalogs. Data used in this research comes from the Sloan Digital Sky Survey and Canada–France–Hawaii Telescope Lensing Survey. It has shown that ConvNets can learn in a completely automated way the features themselves that are relevant for classification, directly from pixel values. With this, the authors of the present work report that their ConvNet model is competitive in accuracy with conventional methods and produces well-calibrated probabilistic classifications necessary for large-scale photometric surveys, such as the Dark Energy Survey and the Large Synoptic Survey Telescope. The investigation underlines that one of the most important issues is the avoidance of overfitting by carefully designing the architecture and tuning hyperparameters. Based on this, all future analyses in astronomy will be efficiently classifiable using the ConvNet framework on this dataset that contains more than 65,000 star and galaxy sources. In their contribution, Kim and Brunner were indicative of the transformation brought forth by deep learning techniques in automating and enhancing the processes for the classification of star-galaxy, allowing sophisticated analyses in astrophysics.

Yu Bai et al. [10] used a machine learning approach based on Random Forests to classify stars, galaxies, and quasars; and to regress stellar effective temperatures. The code combines a large amount of data from LAMOST and SDSS with a nine-color photometric dataset, which includes millions of objects. This gave the RF algorithm classification accuracies of more than 99% for stars and galaxies, and over 94% for QSOs. Furthermore, the standard deviations derived for the stellar effective temperatures were less than 200 K, well within the range of normal spectroscopic methods. The efficiency of machine learning in handling large astronomical data-on which conventional methods normally rely on intricate color cuts or spectral energy distributions-is emphasized in the study. The robustness of RF in handling such large-scale classification tasks is underlined by the authors, who have validated their methodology using blind tests, and hence this work provides an advance toward improved data analysis in astrophysics. That does not only present how machine learning can play a vital role in astronomy but also provides a benchmark for upcoming research studies in order to classify and analyze high-volume astronomical data sets with better efficiency.

In the paper by Jorge de la Calleja and Olac Fuentes[11], the authors have performed classification of galaxies into morphological classes using the technique of machine learning with neural networks and locally weighted regression. Since it improves the accuracy of classification, the authors combined homogeneous ensembles of neural networks by using bagging, and for locally weighted regression they used feature manipulation. The preprocessing of galaxy images was done by rotating, centering, and cropping of images in order to bring standardization in input images. They extracted relevant features using principal component analysis for reducing the dimensionality of features. Their experimental result shows the ensemble of locally weighted regression achieved over 91% accuracy by classifying three types of galaxies, namely Elliptical, Spiral, Irregular. Also, it achieved an accuracy of more than 95% while classifying two types of galaxies: Elliptical and Spiral. This work overcomes some of the challenges of the traditional visual methods, which have proved to be extremely time-consuming and require a lot of expertise, by demonstrating how effective automated methods are in handling large volumes of data coming from digital sky surveys. The results show that machine learning could increase efficiency and accuracy in galaxy classification, therefore opening possibilities for new developments in astronomical data analysis.

P. O. Baqui et al. discuss the application of ML techniques [12] for star-galaxy classification in the miniJPAS survey, which covers the AEGIS field with roughly

1 square degree using 56 narrow-band and 4 broad-band filters. It aims to develop a complementary ML classifier to the traditional methods using a dataset of about 64000 objects crossmatched with reliable classifications drawn from the SDSS and HSC-SSP. They consider six different Machine Learning algorithms, among them RF and ERT, that exhibit outstanding results in terms of classification: the AUC is 0.957 for RF and 0.986 for ERT by using the morphological parameters. The results presented in this work prove that machine learning methods can outperform traditional classifiers at fainter magnitudes, namely for r 21, because the feature characterization has been really complete. Conclusions are drawn by making the value-added catalog and ML models publicly available, having shown the capability of ML in efficiently processing large astronomical datasets that improve the source classification in future surveys like J-PAS.

In the paper by Ninad Shinkar et al.[13], a classification model based on machine learning has been proposed that classifies stars, galaxies, and quasars using data from the Sloan Digital Sky Survey. Algorithms such as RF, SVM, Logistic Regression, Decision Trees, and XGBoost had to be considered in choosing the most appropriate model for correct celestial classification. Their findings also presented that RF had the best accuracy at 97.42%, closely followed by SVM and LR, with accuracies of 95.68% and 94.34%, respectively. The paper goes on to assert that the use of techniques like SMOTE will go a long way in addressing the class balance for improved model performance. Results of this study using a 10,000-observation dataset with 17 features point out the possibility of improving classification efficiency with the use of machine learning techniques by at least an order of magnitude compared with traditional methods. These results represent that not only are there significant reductions in workload from automated techniques in astronomical classification, but even more valued astrophysical research is advanced to comprehend better the expanse of celestial bodies. This, in the end, demonstrates how machine learning can alter astronomy and, in future studies, allow for a more accurate identification and analysis of an astronomical object.

A. O. Clarke, A. M. M. Scaife, R. Greenhalgh, and V. Griguta [14] introduce the development of a machine learning framework to classify 111 million sources that do not have any spectroscopic data from the Sloan Digital Sky Survey. The authors used a dataset of 3.1 million spectroscopically labeled sources in order to train an optimized Random Forest classifier using photometric data from SDSS and the Widefield Infrared Survey Explorer. The derived classification identified 50.4 million galaxies, 2.1 million quasars, and 58.8 million stars, possessing outstanding probabilities of classification-13% of galaxies and 15% of quasars with a probability greater than 0.99. The work addresses topics such as class imbalance and makes use of various methods like Uniform Manifold Approximation and Projection to carry out efficient dimensionality reduction, enabling the visual separation of source types in a two-dimensional space. Discussion of implications for transfer learning on even fainter populations extends the applicability of this model beyond its training set. This extensive classification effort not only helps in efficiently processing really large astronomical data but also provides a significant boost toward knowledge about starry populations and the awe-inspiring capacity of machine learning toward astrophysical research. In all, the results prove the efficiency of automatic classification systems in the process of handling the increased flow of data provided by next-generation telescopes, enabling further research in astronomy.

Chapter 3

Methodology

The methodology for this research is structured to effectively deal with the challenges of classifying stars and galaxies from complex celestial image datasets. It combines advanced image processing techniques, efficient data augmentation and customized deep learning models mainly focusing on convolutional neural networks (CNNs) like ResNet50, which are particularly effective for both binary and multi-class classification tasks. The methodology is subdivided into several sections: Workflow Diagram,Data Collection & Preparation, Model Development, Customizing ResNet50 Architecture, PCA for Structured Data Integration,Hybrid Model Architecture,Training Strategy and Model Evaluation and Metrics each significant for better performance of the model.

3.1 Workflow Diagram

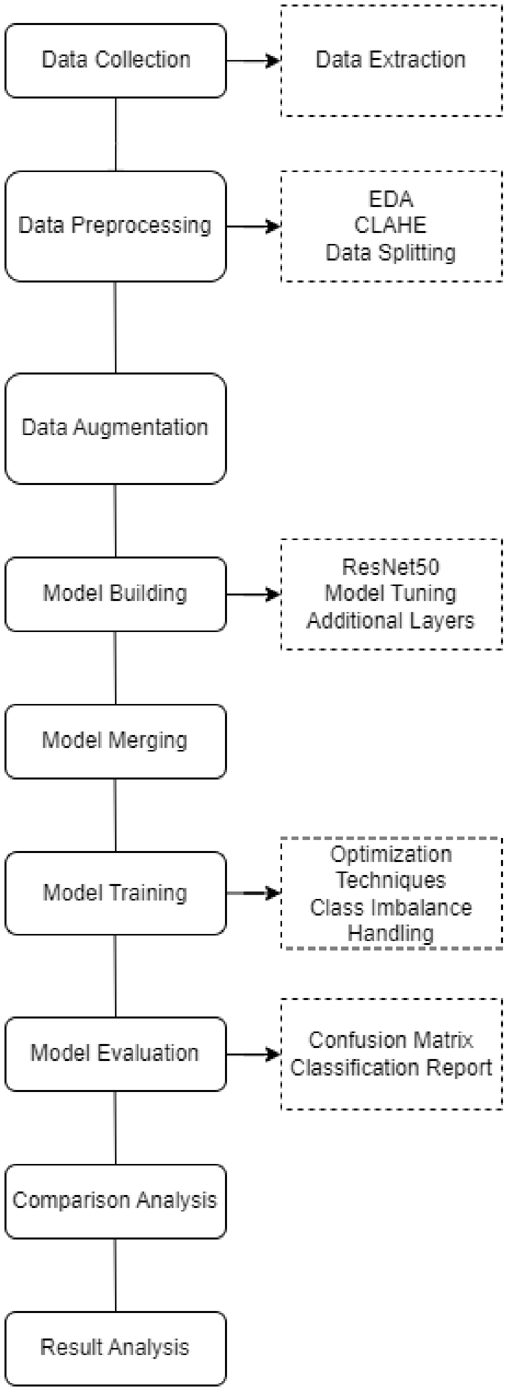

Figure 3.1: Workflow Diagram

3.2 Description of Data

3.2.1 Collection and Preparation of Data

Data Splitting

The dataset is split into an 80% training and 20% testing set. This ensures maximum amount of data for model learning and reserving a significant amount of data for validating the learnt features on unknown dataset.

Sample Image Analysis

Figure 3.2: Sample Galaxy Images

Galaxies: Sample images of galaxy dataset is displayed here which contains varying brightness levels and different structures highlighting data diversity.

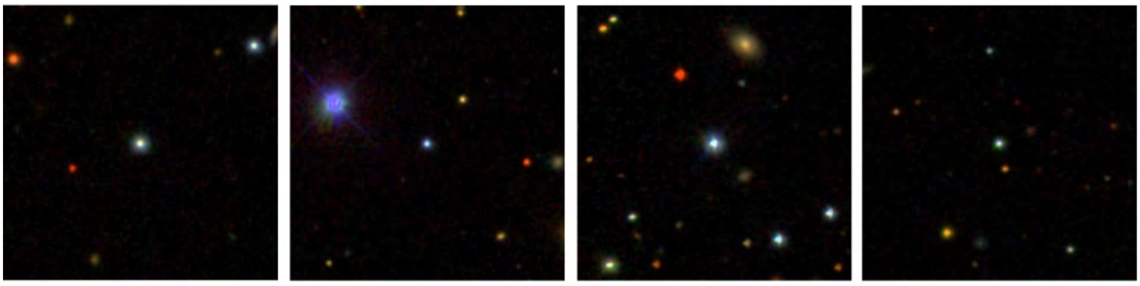

Figure 3.3: Sample Star Images

Stars: Images of stars focus on exhibiting their typical point-like and bright center characteristics, which differentiate them from galaxies in astronomical observations.

Data Preprocessing

- **CLAHE (Contrast-Limited Adaptive Histogram Equalization):** This technique is used to enhance the contrast of the images, making it easier for the model to detect subtle differences in celestial object features.
- **Normalization:** All input data are scaled to a uniform range, typically 0 to 1, which helps in stabilizing the neural network's gradient descent optimization process.

- **Data Augmentation:** To make the model robust against variations in new data, techniques such as random rotations (up to 20 degrees), shifts (up to 10% of the image size), zooms (up to 10%), and brightness modifications are applied. These augmentations help simulate different observational conditions and instrumental variations.

Model Development

Layer (type)	Output Shape
input_layer_1 (InputLayer)	(None, 2052)
dense (Dense)	(None, 256)
batch_normalization_1 (BatchNormalization)	(None, 256)
dropout (Dropout)	(None, 256)
dense_1 (Dense)	(None, 128)
batch_normalization_2 (BatchNormalization)	(None, 128)
dropout_1 (Dropout)	(None, 128)
dense_2 (Dense)	(None, 64)
batch_normalization_3 (BatchNormalization)	(None, 64)
dropout_2 (Dropout)	(None, 64)
dense_3 (Dense)	(None, 2)

Figure 3.4: Custom ResNet50 Model Summary

The model development phase focuses on constructing a robust deep learning framework that can efficiently classify images of celestial objects. Here, we elaborate on the customization of the ResNet50 architecture and the integration of structured data to enhance model performance.

Customizing ResNet50 Architecture

Image Feature Extraction: Leveraging the pre-trained ResNet50 architecture, which is originally designed for 1,000 object categories, we modify it for astronomical purposes. The ResNet50 uses residual blocks, allowing it to learn deeper features without being affected by gradient problem. These blocks are important for maintaining strong gradient flow which helps the model to learn from a very deep network structure.

Batch Normalization and Activation: Each convolutional layer has a batch normalization section, which normalizes the activations from the previous layer, speeding up convergence and improving overall model's performance. ReLU (Rectified Linear Unit) activation functions are used into the mode which helps in adding non linearity to the models, helping it to learn complex patterns among variables in the data.
Global Average Pooling (GAP): Instead of flattening the feature maps and then connecting them to dense layers, GAP reduces each feature map into a single number by averaging out the values. This decreases the model parameters and subsequently helps in minimizing overfitting.

PCA for Structured Data Integration

Data Preprocessing: Preprocessing under structured data involves celestial coordinates and photometric measurements prior to integration. This involves normalizing the data to ensure that all input features contribute equally to the analysis, preventing features with larger scales from dominating the learning process.
Dimensionality Reduction: PCA is applied to reduce the dimensionality of the structured data while retaining 95% of its variance. This step simplifies the data, reducing training time and improving model generalization by focusing on the most informative features.

Hybrid Model Architecture

Combining Features: The reduced PCA features from the structured data are concatenated with the image features extracted by the ResNet50 layers. This combined feature set provides a richer representation of each celestial object, enhancing the model's ability to distinguish between stars and galaxies.
Dense Layers and Regularization: After concatenation, the features are passed through several fully connected (dense) layers. Each dense layer is followed by batch normalization and dropout, where dropout randomly sets a fraction of input units to zero during training, helping to prevent overfitting.
Output Layer: The final layer of the model is a softmax layer that classifies the inputs into two categories: stars and galaxies. This layer outputs the probabilities for each class, facilitating the classification decision.

Training Strategy

Data Augmentation: To enhance model robustness and ensure better generalization, we implemented a data augmentation strategy on the image data. This included applying minor rotations up to 20 degrees, horizontal and vertical shifts of up to 10%, zoom operations up to 10%, and brightness variations to mimic different observational conditions. These transformations help the model learn from a more diversified dataset, simulating a variety of astronomical viewing conditions.
Class Imbalance Handling: The dataset exhibited a significant imbalance between the classes representing stars and galaxies. To address this, we computed class weights from the training data distribution, which were then applied to the loss function during training. This approach helps mitigate the model's bias toward the more frequently occurring class by adjusting loss contribution from each class.

Training Environment and Optimization Techniques: The training was conducted on Google Colab, utilizing a Tesla T4 GPU, which provided the necessary computational power to process large datasets efficiently. We used the Adam optimizer for its adaptive learning rate capabilities, setting an initial learning rate of 1e-4. To combat overfitting and ensure the model was not just memorizing the training data, we implemented early stopping, which halts training if the validation loss does not improve after several epochs.This was further complemented by a learning rate scheduler that reduced the learning rate by a factor of 0.05; if no improvement in validation loss was observed for a couple of epochs consecutively, it allowed more fine-tuned adjustments towards the later part of training. Additionally, the model was set to train a total of 5 epochs, while these callbacks made sure that optimal training dynamics would not come at an excessively high computational cost.

Model Evaluation and Metrics

Comprehensive Metrics: It includes quantifying model performance with accuracy, precision, recall, and F1 scores, with even more granular diagnostics such as confusion matrices and classification reports which give much more insight into the performance class by class.
Visualization Techniques: Activation maps visually represent which areas of the images most influence the model's decisions, while training and validation curves monitor model performance across training epochs, offering insights into learning stability and effectiveness.

Optimization for Future Applications:

Scalability and Adaptability: The scalability and adaptability of the proposed methodology and architecture of the model are such that they can scale up in case of bigger datasets or any other type of astroinformatics classification problems, hence this research is capable of doing further application in future multi-class classification challenges.[15]

3.2.2 Dataset Overview and Image Characteristics

The data used in this research are astronomical images from some large-scale sky survey projects, notably the Sloan Digital Sky Survey[16]. Such surveys have mapped a wide range of celestial objects-stars, galaxies, quasars, and many other astrophysical objects-during the last couple of decades. The data in this project are drawn to suit specifically binary classification of two most primary objects in the study of cosmology and astrophysics: stars and galaxies. As both capture enormous quantity and variety, the dataset was an opportunity and a challenge in itself. Deep learning models can learn complex representations from large volumes of images, but rigorous preprocessing is needed to make sure the quality of the data is good enough to train high-performance machine learning models. Additional information on the data is gathered by performing Exploratory Data Analysis on the SDSS dataset used here.

Source of the Data

These are primarily sourced from the SDSS, one of the largest astronomical surveys undertaken until date, which mapped one-third of the sky and obtained images for hundreds of millions of objects. These images have a lot of rich metadata associated with them, like right ascension, declination, redshift, and spectral data from which the objects will be classified into various astronomical classes.
The images for this project were extracted from the website of SDSS through Python scripts. In particular, the images were downloaded by linking them to another CSV file that contained key features in the SDSS database, such as object coordinates and their classifications. This approach is automated and therefore ensures correct mapping of image files to the features of the dataset that need further analysis.
In this work, the dataset filtered from this research contains only images of star and galaxy categories. This is a binary problem that would enable focusing on the saliently different visual features presented by these two kinds of objects, which is so important for the study of the large-scale structure of the universe and stellar evolution.

Image Characteristics

The specific set of images each represents a star or a galaxy, as captured by different space- or ground-based telescopes. The general characteristics of the images are as follows:

– **Format of the Images:** The images are provided in JPEG or PNG formats, hence being compatible with most of the image processing and machine learning libraries[17]. These formats ensure that file sizes are manageable, even when working with thousands of images. [18]

– **Image Resolution:** The original images vary in resolution, depending on the specific telescope or instrument used to capture them. For uniformity and computational efficiency, all images were resized to a standard resolution of 256x256 pixels. This resolution balances the retention of enough visual detail for classification while keeping the computational requirements reasonable.

– **Color Channels:** Most of the images are color images captured in multiple bands (typically five: u, g, r, i, z, representing ultraviolet, green, red, near-infrared, and infrared, respectively). For simplicity, the images were converted to three-channel RGB format, maintaining critical visual information while reducing the complexity of multi-band analysis.

– **Variability in Brightness and Contrast:** Images within this dataset differ strongly in brightness and contrast. While some images can show very faded and distant stars or galaxies, other images show bright, very close-by objects. This complicates the process of classification by models in classification due to the fact that they require image preprocessing techniques such as enhancement[19] to regularize the visual quality of images.

3.3 Data Preparation

This dataset used in the study contains both structured information in the form of CSV data and unstructured information pertaining to images from the SDSS. Each row also contains the most important metadata, such as the celestial coordinates RA and DEC; photometric information in multiple filters like u, g, r, i, and z; and redshift, a very rich set of features for accurate classification.

CSV Data Preparation of CSV Data

For the quality and relevance of the data, the CSVs were put through an exhaustive pre-processing routine. This has included removal of columns like 'objid', 'run', and 'fiberid', which are pretty less useful for classification tasks. It also includes sorting appropriately the data to match image data. 'Class' column omitted to utilize model properly. The rest of the features are then standardized using the MinMaxScaler, scaling them to one uniform scale that increases the sensitivity of the model to subtle variations in the input data. Image and CSV Data:

Image and CSV Data:

Merging The integration of image and CSV data was an important step in preparing a complete dataset for training the deep learning model. Each image file, which corresponds to a celestial object, was matched to its metadata by aligning the 'sdss_number' from the CSV file with the image filename. It aligns such that every input instance to the model would consist of both pixel data and enriched metadata, providing a comprehensive view to effectively discern between stars and galaxies. This preprocessed combined dataset then leverages the strengths of both image-based and feature-based machine learning to make the model more robust against any variation in image quality and more insightful by adding metadata information.

Labeling and Classes

The dataset consists of two major classes:

- **Stars (Class 0):** Luminous spheres of plasma held together by their own gravity. The stars usually appear more point-like with variable brightness. The dataset includes a range of star types, adding complexity to distinguishing them from galaxies.
- **Galaxies (Class 1):** Vast collections of stars, gas, and dark matter bound by gravity. Galaxies exhibit diverse shapes, such as spiral arms, elliptical halos, or irregular forms. The dataset includes various galaxy morphologies, from spiral galaxies to elliptical ones.

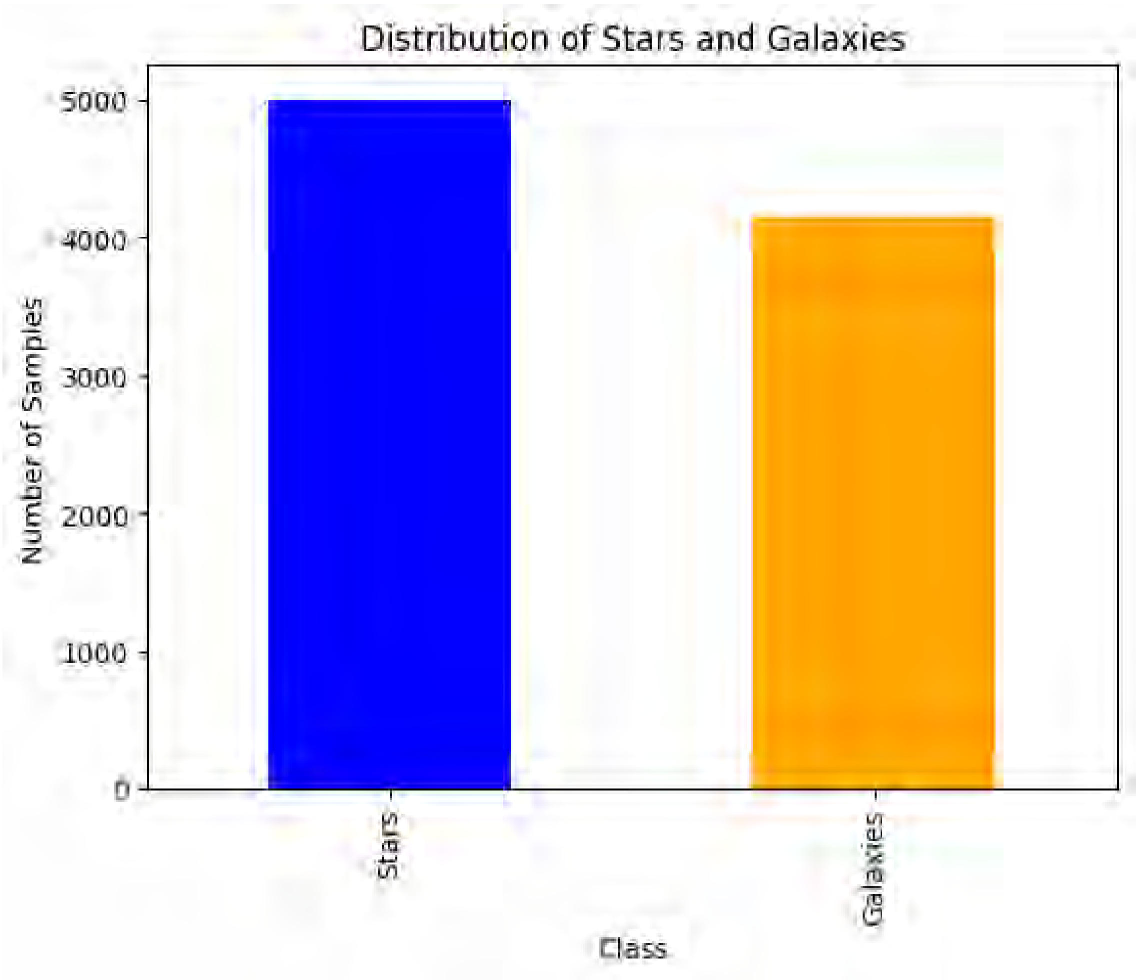

Figure 3.5: Class Distribution of Stars and Galaxies

The class distribution is slightly imbalanced, with stars being more prevalent than galaxies. This imbalance presents a challenge for training machine learning models, as they may become biased toward the majority class (stars). Class weighting and data augmentation techniques were employed during training to address this issue and ensure the model can classify galaxies with similar accuracy to stars.

Metadata

In addition to the image data, each image is accompanied by rich metadata provided by the SDSS database. This metadata includes important astronomical attributes for each object:

- **Right Ascension (RA):** The celestial coordinate that defines the object's position along the celestial equator, analogous to longitude on Earth.
- **Declination (DEC):** The celestial coordinate that defines the object's position north or south of the celestial equator, analogous to latitude on Earth.
- **Redshift (z):** A measure of how much the object's light has shifted to the red end of the spectrum due to the expansion of the universe. This is especially

important in the case of galaxies since redshift can give information on the distance and velocity of objects.

– **Spectral Information:** In some cases, certain objects have additional spectral data which describes the intensity of the light emitted at different wavelengths. While not directly useful for classifying images[20], this information could be useful for further studies.

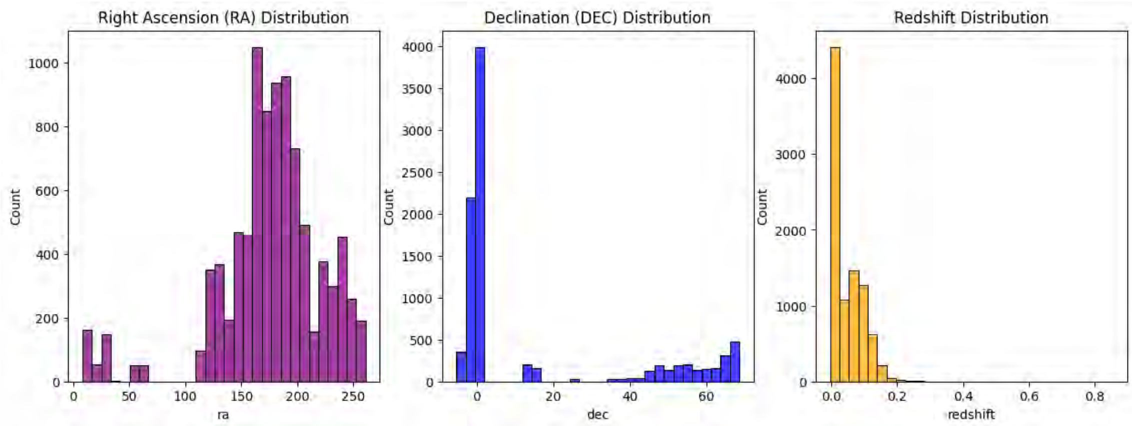

Figure 3.6: Distributions of RA, DEC, and Redshift from the SDSS Star-Galaxy Dataset

This plots the distributions of three key metadata features from the SDSS star-galaxy image dataset in CSV format. These features include RA, DEC, and Redshift-three important astronomical parameters used in the identification of celestial objects.
Right Ascension (RA) Distribution (Left Plot):
X-axis (ra): This is the Right Ascension, the celestial equivalent of longitude in defining an object's position in the sky.
Y-axis (Count): Shows how frequently different RA values appear in the dataset.
Shape: The RA distribution takes an approximately normal distribution centered around the 150-200 degree range. That indicates the majority of the celestial objects in this dataset are congregated within that particular RA range.
Declination (DEC) Distribution (Middle Plot):
X-axis (dec): represents the declination, which is the celestial term for latitude. This is a measure of how far something is north or south of the celestial equator.
Y-axis (Count): The count of the number of celestial objects for different declination values.
Shape: The distribution of DEC is highly skewed, lying mostly below 0, suggesting that most objects are at or near the celestial equator, with a lesser spread of the objects at higher declination values either north or south.
Redshift Distribution (Right Plot):

X-axis (redshift): This is a dimensionless quantity representing the Redshift, which refers to the amount of stretching the light of an astronomical object has undergone by the expansion of the universe. Higher redshift values typically indicate objects that are farther away and moving away faster

Y-axis (Count): Shows the count of objects at different redshift values.
Shape: The redshift distribution is heavily skewed to the left, with a peak near 0.0 and a sharp drop-off as redshift increases. This indicates that most objects in the dataset have low redshift values, implying that they are relatively close to Earth.

The RA and DEC distributions give insights into the spatial coverage of the celestial objects in this dataset.
The redshift distribution suggests that the dataset is dominated by nearby celestial objects, as very few objects have high redshift values.

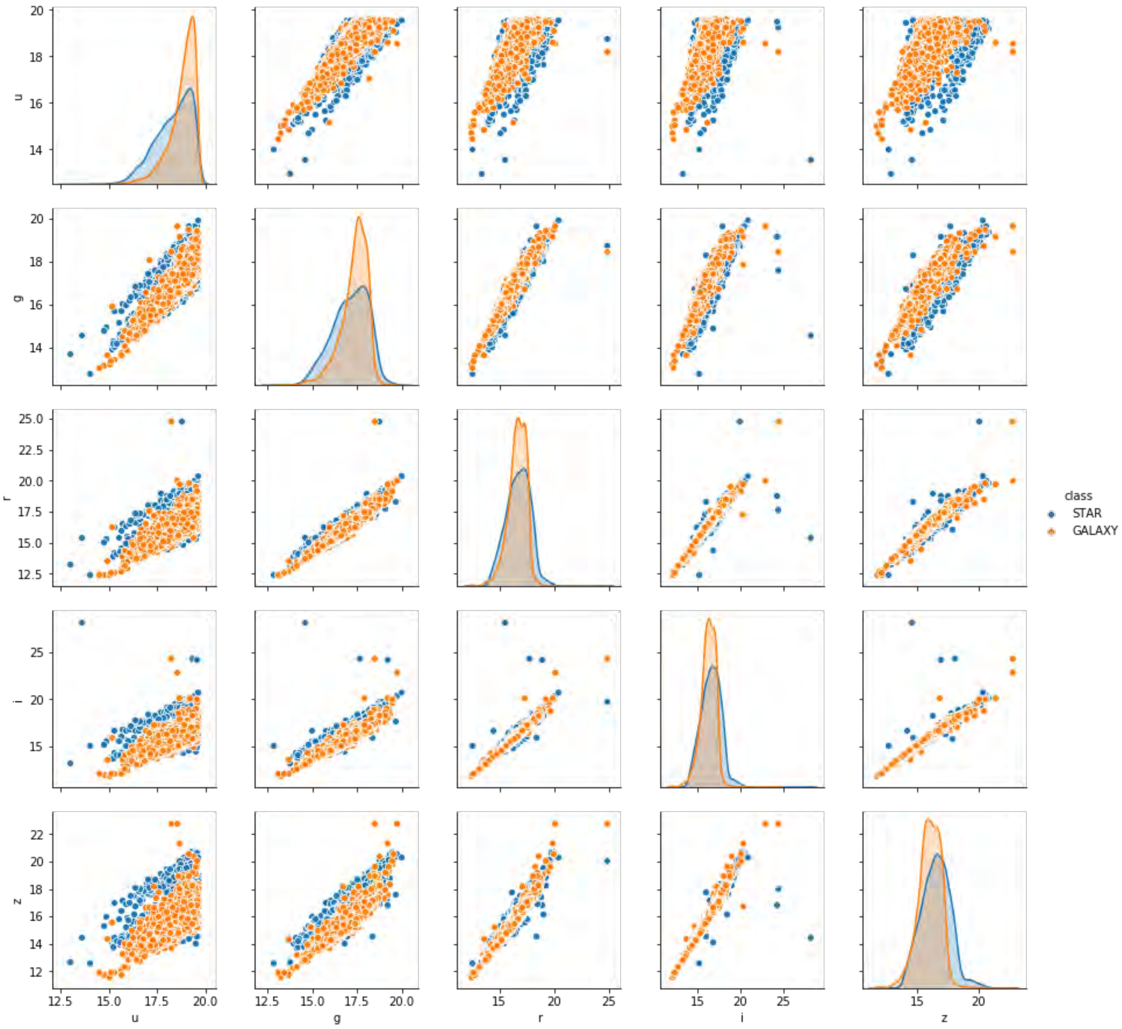

Figure 3.7: Photometric Analysis of u, g, r, i, z Magnitudes: Scatter Plots and Histograms

In the presented visualization, pairwise scatter plots and histograms depict the distribution of u, g, r, i, z magnitudes across classified stars and galaxies. The data extracted and visualized from the Sloan Digital Sky Survey (SDSS) offers critical insights into the photometric characteristics differentiating celestial objects.

Pairwise Scatter Plots: The pairwise scatter plots between various photometric magnitudes (u-g, u-r, u-i, u-z, g-r, g-i, g-z, r-i, r-z, i-z) reveal distinct clustering patterns of stars and galaxies, indicating differential light absorption and emission profiles in these objects. Galaxies tend to cluster differently from stars, often showing more extended and dispersed distributions. This is indicative of broader spectral energy distributions in galaxies compared to stars. Moreover, the dispersion and concentration of points in these plots reflect changes in observed magnitudes as a function of galactic structure and star formation history.
Histograms and Density Plots: The diagonal plots are histograms with kernel density estimates overlaid to display the univariate analysis for each magnitude. Such plots explicitly present the very different modes of the distribution for stars and galaxies, bringing out the skewness and kurtosis inherent in data of this nature. Take, for example, the r-band magnitude, since it has a clear peak for stars due to their sensitivities toward stellar temperatures. Contrarily, galaxies show a wider spread in the same band, thereby indicating a composite light profile from different stellar populations. Implications for Classification: The patterns seen in the scatter plots and histograms form the backbone of developing the classification algorithms. Characteristics such as the spread in the i-z plane and the distinct peaks in the r-band histograms are really useful in distinguishing between stars and galaxies. These same visual cues are in harmony with theoretical expectations, whereby galaxies, being multiple star systems and interstellar material, give more complex and variable photometric signatures compared to single stars.
Statistical Correlation: The visualization helps in taking a deeper look at the different magnitudes and their correlation, which is important for the selection of features in machine learning models focused on classification of the objects that are observed in space.The degree of linear association visible in the pairwise plots suggests that some magnitudes can serve as proxies for others, potentially reducing the dimensionality of the problem without significant loss of information.
This metadata enhances the richness of the dataset, offering further avenues for astronomical analysis, such as studying the spatial distribution of stars and galaxies or correlating redshift with galaxy morphology.

Dataset Size and Distribution

The dataset contains approximately 9,150 images, divided into training and testing sets as follows:

- Training Set (80%): Approximately 7,319 images of stars and galaxies were used for training the deep learning model. This set was augmented with transformations (rotation, zooming, shifting, etc.) to increase the dataset's diversity.

- Testing Set (20%): A subset of 1,831 images was reserved for testing the model's performance on unseen data. This section was reserved and unchanged which ensures that the model's evaluation was based on real, unaltered images.

Libraries Used

In the development of our model for classifying celestial objects using space telescope images, we employed a variety of libraries and tools that facilitated our data processing, model building, and evaluation tasks:

Data Handling and Preprocessing:

- **Pandas:** This is used for manipulating data and reading/writing to the CSV format.

- **NumPy:** It is used for the implementation of numerical operations on arrays and matrices, something so basic when handling and processing data.

- **scikit-learn:** This library provides the necessary data pre-processing functions utilized in this work. This range includes the MinMaxScaler to scale the features and PCA for dimensional reduction.

Model Development and Training:

- **TensorFlow and Keras:** These libraries provided the core for our model development and training. TensorFlow[21] acted as the backend framework, while Keras made things easier to implement in terms of building neural networks.

- **ResNet50:** Although this pre-trained model was taken from Keras applications, some changes were done to the top layers. Retraining the model allowed it to classify between stars and galaxies based on their images.

- **ImageDataGenerator:** This is used from Keras for real-time data augmentation. Real-time data augmentation is one of the most critical tasks for robust model development to simulate varied conditions of the astronomical images.

Optimization and Regularization Techniques:

- **Adam Optimizer:** This optimizer is chosen as it is efficient while converging faster by adapting the learning rate throughout training.

- **Dropout and Batch Normalization:** We have incorporated dropout and batch normalization into our network architecture in order to prevent overfitting and improve model generalization.

- **L2 Regularization:** This regularizes the dense layers in our network and adds more constraints to model complexity to avoid overfitting.

Model Evaluation:

- **EarlyStopping, ReduceLROnPlateau:** These two Keras callbacks were used for in-model performance monitoring, allowing the model to perform early stopping and reduction of learning rate on a plateau, among other things, if the loss validation stopped improving.

- **Compute Class Weight:** Scikit-learn function to handle class imbalance by adjusting the weight inversely proportional to class frequencies in the input data.

Combined Data for Model Training

Information used in this paper improves by combining image data and related metadata features. Metadata features like RA, DEC, u, g, i, r, z, and redshift were extracted from another CSV file in Python. The above features have gone through scaling with the ‘MinMaxScaler‘ to make their values consistent in a similar range. Then, after normalization across different units and scales, PCA was implemented which reduced the dimensions with 95% preserved variance, hence increasing computational efficiency and model simplicity.
The ResNet50 model, pre-trained on ImageNet[22], was adapted to extract deep features from the images. These images were pre-processed through a custom data pipeline that included rescaling, different augmentation techniques such as rotation, shift, zoom, and brightness, and finally resizing to 256x256 pixels. The image features are then concatenated with the PCA-reduced CSV metadata features to provide a rich and combined dataset that can enable the model to learn from both the visual and tabular aspects of the data. The features extracted from the images were combined with the normalized and reduced metadata for the final combined dataset, and all combined data was aligned properly by sorting the rows in ascending order on the ’sdss_number’ field. In this way, a model can capitalize on both the image features and the tabular ones and, as a result, make the model improve its ability to classify star-galaxy classes more accurately.

3.4 Data Pre-Processing

Data preprocessing is an indispensable ingredient in all machine learning pipelines, let alone when working with high-dimensional image data. After all, this directly impacts the quality of the input into your model and in turn affects its capability to learn useful features. Consequently, several of the key steps considered in preprocessing were to improve quality, standardize the dataset, and prepare the images for good usage in a deep learning model. These are pre-processing steps, specifically developed to handle problems that have been developed based on variability in image quality, class imbalance, and artifacts of the complexity of celestial objects. The following describes the preprocessing techniques in detail for this study.

3.4.1 Data Extraction and Organization

Organization of the dataset was done first in the chain of preprocessing. The images of stars and galaxies were extracted from their compressed archives into two separate directories for each class: star and galaxy classes. This further increased the separation between the classes, hence making handling easier both in the training and validation and test sets. After that, the whole dataset is divided into training and test subsets in a ratio of 80:20, ensuring that the test. Thus, the set remained intact during the training so that it could give an unbiased measure of the performance of the model.

3.4.2 Image Enhancement Using Contrast-Limited Adaptive Histogram Equalization (CLAHE)

Many of the astronomical images, especially those taken from ground-based telescopes, inherently possess very poor contrast due to atmospheric interference and other light-related conditions that generally deteriorate the quality of the image. In this respect, Contrast-Limited Adaptive Histogram Equalization was used on each image in the dataset. CLAHE is one of the more advanced enhancement tools which seeks an improvement of the local contrast without amplifying noise and therefore is particularly fitted for such types of images in astronomy where details are most of the time very minute, hence rather weak.

a) **Explanation**
This is the histogram equalization technique applied to small, non-overlapping tiles (local regions) of the image. This technique redistributes the pixel intensity values within each local region in such a way that the contrast becomes optimally stretched locally without over-amplifying noise. The "contrast-limited" aspect of the term makes sure that the enhancement in contrast is enraged but controlled to avoid artifacts that may arise from over-enhancement.

b) **Use**
What was done for each image in the dataset includes the following:

- LAB Color Space Conversion: Images were first pre-processed by converting them from RGB to the LAB color space; the lightness of an image refers to the L channel, which represents the intensity or brightness of the image. The algorithm went through image processing that modified the L channel, keeping the color information-already represented by the A and B channels.
- CLAHE Processing: It was performed with the L channel in a 3x3 tile grid with a clip limit of 5.0. This ensured enhancements of local contrast in each small region of the image to make ghostly details such as galaxy arms or stellar edges more apparent.
- Reconversion to RGB: Once CLAHE was applied to the L-channel, conversion to the RGB color space produced an enhanced image in better contrast and clarity.

This technique greatly optimized the visibility of fine details in the images, such as the spiral arms of the galaxies and the brightness variations of the stars, which distinguishing features were important for the deep learning model.

3.4.3 Image Resizing

Original images were of variable size, as it depended on a particular telescope and the object under observation. All the images were uniformly brought to a standard resolution of 256x256 pixels to decrease the computational overhead. Such a size was chosen to balance the need to retain important visual details with the need to manage memory and processing resources efficiently.

a) **Preserving the Aspect Ratio**
The ROIs were resized in a way that the aspect ratio of each image was preserved. This is important to maintain geometric integrity for celestial objects, such as galaxy shapes or the structure of star clusters, for proper classification.

b) **Bilinear Interpolation**
It explicitly uses bilinear interpolation to resize an image for downsizing with minimal loss of detail. For the computation of a pixel value in the resized image, bilinear interpolation considers the nearest four pixels of the image. This yields a smooth result and does not introduce significant artifacts.

3.4.4 Image Normalization

Image normalization was done to standardize the pixel values within the dataset. Since the pixel values in raw images vary from 0 to 255, normalization was used to scale these values in the range between 0 and 1 for consistency and to help in the ease of training of the deep learning model. This allows the model to converge faster during training and helps the gradients be more stable during the process of backpropagation.

a) **Scaling of Pixel Value**
To normalize the values of the pixels in the images to fall within a range from 0 to 1, each value was divided by 255. Normalization here is a very important step because it normalizes the range of input values over all images so that no one image dominates the training process because of large values of pixel intensity.

b) **Model Training Effect**
Firstly, normalization prevents the explosion or vanishing of gradients during backpropagation; hence, convergence is much faster and more stable. Moreover, the model would learn features much more efficiently from normalized inputs rather than being biased to pixel intensity, letting other useful spatial information go unnoticed.

3.4.5 Data Augmentation

Data augmentation was applied to the training set to artificially increase Data augmentation was applied to the training set with the aim of artificially increasing both size and variability of this dataset. This augmentation helps in creating new training examples through applying various transformations on the original images, hence effectively reducing overfitting and enabling the model to generalize better on unseen data. The following augmentation techniques were employed:

a) **Random Rotation**
Random rotations were applied to simulate different orientations of celestial objects.For example, galaxies may appear at various angles depending on their position concerning Earth. One way this can be done is for example, if rotated images by random angles within a certain range, say from -20 to +20 degrees which is used in our model, so the model could learn rotational invariant features of objects.

b) **Horizontal and Vertical Shifts**
Random horizontal and vertical shifts were applied to simulate changes in the position of objects within the image frame. This helped the model learn to recognize objects regardless of their position within the image, making it more robust to different framing and centering of celestial objects.

c) **Zooming**
Random zooming was employed to simulate changes in the apparent size of celestial objects, which occur naturally due to their varying distances from Earth. By zooming in and out (within a specified range, such as 0.1x), the model learned to classify objects regardless of their size, improving its ability to generalize across different scales.

d) **Horizontal Flipping**
Horizontal flips were applied to the images to simulate mirror images of celestial objects. This is particularly useful for galaxies, as they can appear in various mirrored orientations. Flipping increased the variability of the training set without requiring additional labeled images.

e) **Brightness Adjustments**
Random brightness adjustments were made to simulate different lighting conditions in the sky. Celestial objects can appear with varying brightness levels depending on atmospheric conditions or their distance from Earth. By randomly adjusting the brightness (within a range of [0.9, 1.1]), the model became more robust to variations in exposure and light levels, improving its performance on images with poor illumination.

3.4.6 Class Label Encoding

Since this research focused on binary classification (stars versus galaxies), the class labels were encoded numerically:

- Stars were labeled as 0.
- Galaxies were labeled as 1.

This encoding was essential for transforming the categorical labels into a format suitable for the model's output layer, which used softmax activation to predict the probabilities of each class.

a) **Binary Encoding for Efficient Classification**
By encoding the labels as 0 and 1, the problem was framed as a binary classification task, where the model's goal was to predict whether an image represented a star or a galaxy. This simplified the classification process and ensured that the model's loss function (categorical cross-entropy) could be applied effectively.

b) **Handling Class Imbalance**
The dataset exhibited some class imbalance, with stars being more prevalent than galaxies. To prevent the model from becoming biased toward the majority class, class weights were computed and applied during training. These

weights ensured that the minority class (galaxies) received appropriate attention during the training process, improving the model's overall ability to classify both classes accurately.

3.4.7 Splitting the Dataset and Generating CSV Files

The data was then split into a training subset and a test subset to ensure fairness in model evaluation. Additionally, the feature-normalized CSV data with celestial coordinates was combined with the extracted features of the image to form a complete dataset for training purposes.

3.4.8 Combining Dataset

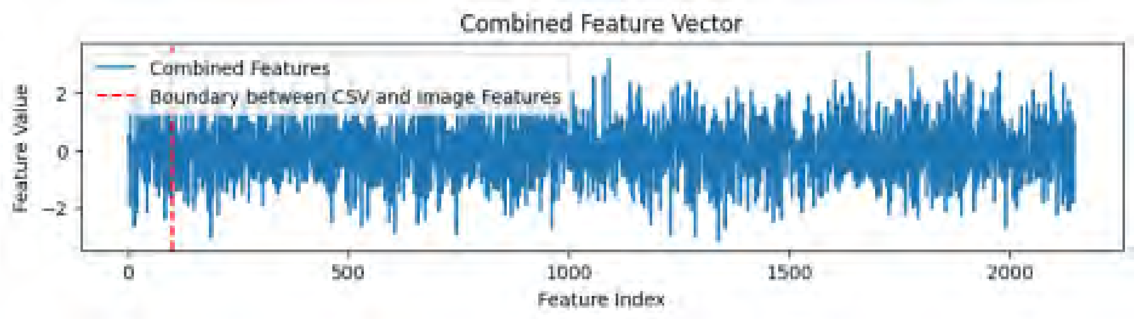

Figure 3.8: Visualization of the Combined Feature Vector

The feature vector in the figure combines both the photometric CSV features and those derived from an image, under one umbrella to form a unified model input. The left portion of the vector, as separated by the red dotted line, represents the PCA-reduced CSV features. These encompass major astrophysical measures such as the coordinates of the celestial objects and the photometric data. The features on the right side of the red dashed line were obtained from the astronomical image by a deep convolutional neural network. This visualization underlines the hybrid nature of the input data: the more classic tabular data is merged with the result of modern image processing in order to enhance both the precision and robustness in the classification of celestial objects. This will enable the model to draw upon a more complete suite of attributes for each input sample, hence allowing a much richer understanding and identification of the various celestial objects that are being modeled. The above-mentioned method can mitigate the drawbacks that arise when a model is working on purely image-based data, which cannot generalize well across all different kinds of astronomical observations.

Chapter 4

Analysis

4.1 Performance Measure

The research on the classification of stars and galaxies explained the adoption of a set of machine learning algorithms with image preprocessing techniques, proving that the methodology adopted gave high accuracy in classifying stars or galaxies. It was trained on more than 50 epochs and used early stopping to avoid overfitting, with learning rate reduction for the optimization in training.

4.1.1 Evaluation Metrics

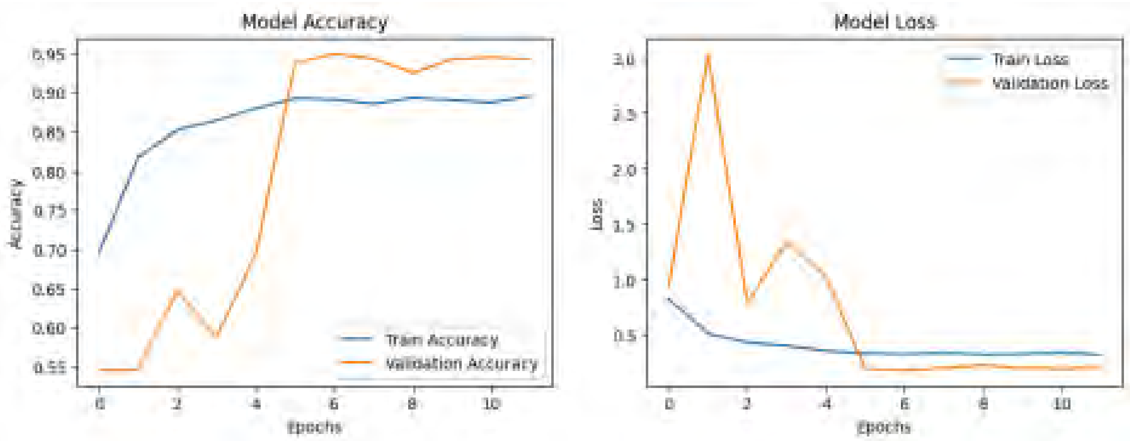

Figure 4.1: Accuracy and Loss Curves of Custom ResNet50

In this case, the model achieved an accuracy of 95%, meaning this particular fine-tuned ResNet50 model did a proper job in classifying stars and galaxies. The test loss of 0.20 further strengthens this evidence of the robustness of the model.

The classification report detailed precision, recall, and F1-score for both classes. Precision was slightly higher for galaxies, 96% for galaxies, and 94% for stars. Both recall and F1-score metrics were also high, which underlined that the performance for both categories was rather well balanced.

	Precision	Recall	f1-score	Support
Galaxy	0.96	0.95	0.95	1000
Star	0.94	0.95	0.95	831
Accuracy				**0.95**
Macro avg	0.95	0.95	0.95	1831
Weighted avg	0.95	0.95	0.95	1831

Table 4.1: Classification Report

4.1.2 Confusion Matrix

Figure 4.2: Comparison of performance metrics across traditional ML models

The confusion matrix allows looking more precisely at the performance of the model. Generally, the matrix had an overwhelming proper classification of test images, with just a few percent of misclassifications of images. Most strikingly, the stars sometimes misclassified galaxies and vice versa. However, in general, the misclassification rates were not profound. From the visualization of the confusion matrix, there is a reliability of the model on making more accurate predictions about galaxies.

4.1.3 Model Summary

The training and validation accuracies continued to rise through the course of the epochs, having a minor divergence between the two curves; this points to the fact

that overfitting was low. This has been achieved by the use of data augmentation and class weighting. Additionally, training and validation loss metrics showed smooth convergence, further supporting the effectiveness of the training strategy.
The visualizations of accuracy and loss during training provided clear evidence of the model's learning process, showing consistent improvement without abrupt fluctuations.

4.2 Comparison Analysis

4.2.1 Accuracy Comparison

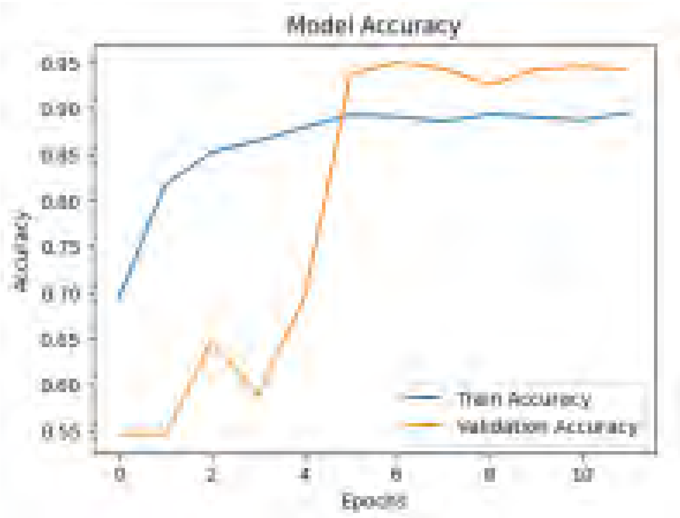

Figure 4.3: Custom ResNet50

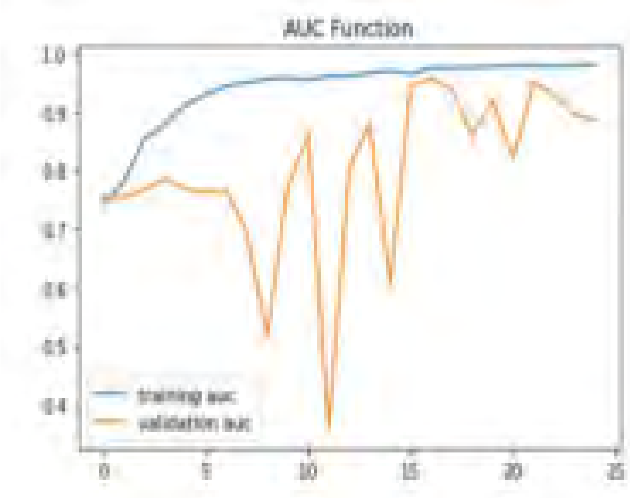

Figure 4.4: Custom CNN

Our custom ResNet50 model achieved 95% accuracy on the test dataset using a fine-tuned ResNet50 model and modified dataset combined with image preprocessing techniques.
The model we are comparing to achieved 92.44% accuracy using a proposed CNN

model with 8 convolutional layers and three fully connected layers for classification on the ARIES dataset.

4.2.2 Model Architecture Comparison

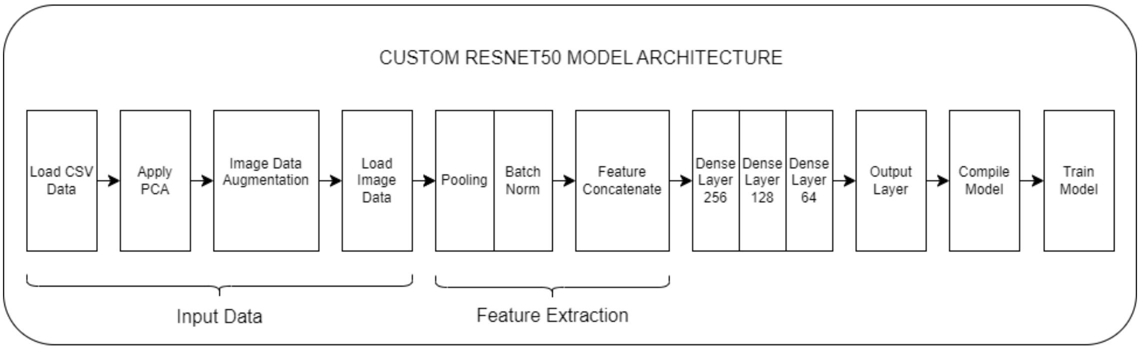

Figure 4.5: Custom ResNet50 Architecture

We used a pre-trained ResNet50 model with transfer learning. The final layers were fine-tuned, and additional dense layers were added for classification. This model benefits from the depth of ResNet50 and transfer learning from large-scale pre-trained networks manrlike ImageNet.

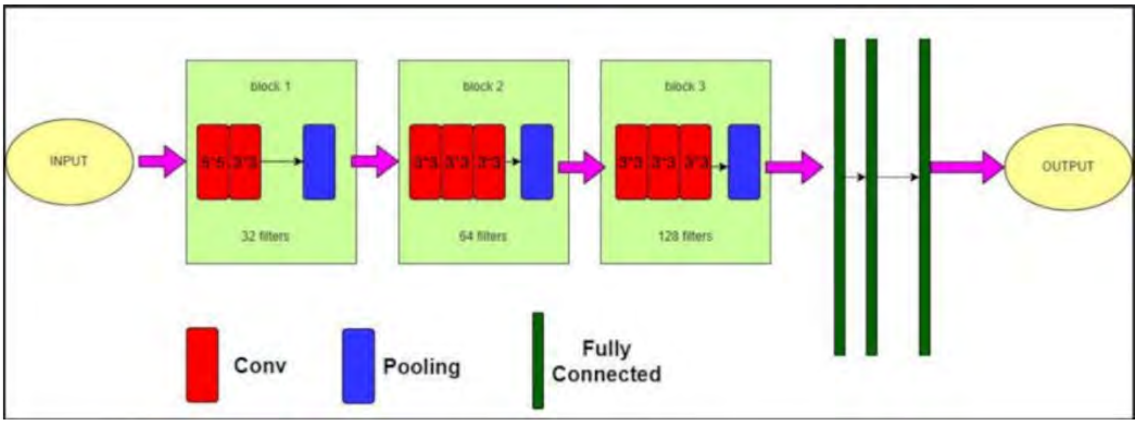

Figure 4.6: CNN Architecture

A custom CNN architecture consisting of 8 convolutional layers followed by 3 fully connected layers. Each block in the CNN progressively extracted and refined features before the final classification

4.2.3 Classification Report Comparison with traditional machine learning algorithms

	ResNet50	**DTC**	**NB**	**LR**
Accuracy	94%	71%	68%	74%
Recall	94%	81%	83%	81%
Precision	94%	80%	74%	88%
F1-score	94%	81%	78%	84%

Table 4.2: Comparison of Performance Metrics Across Different Models

This table compares the performance of traditional Machine Learning algorithms like Decision Tree Classifier, Naive Bayes and Logistic Regression with our custom deep learning ResNet50 model by using accuracy, recall, precision and F1-score as the comparison parameters. We can clearly see that our model is much more efficient and well performing than most other traditional ML models. [23]

4.2.4 Enhancement of Model Performance through Integrated Feature Sets

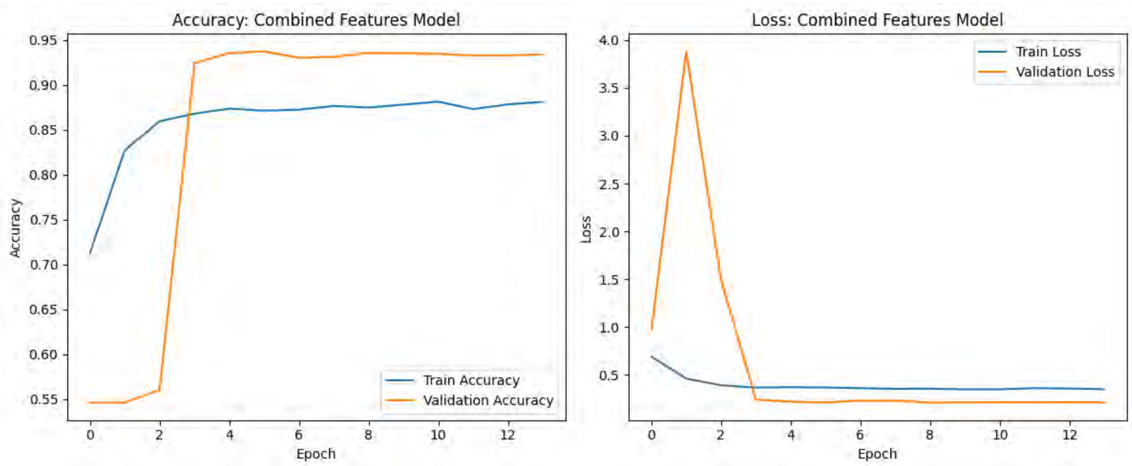

Figure 4.7: Combined Features Model Accuracy and Loss Curves

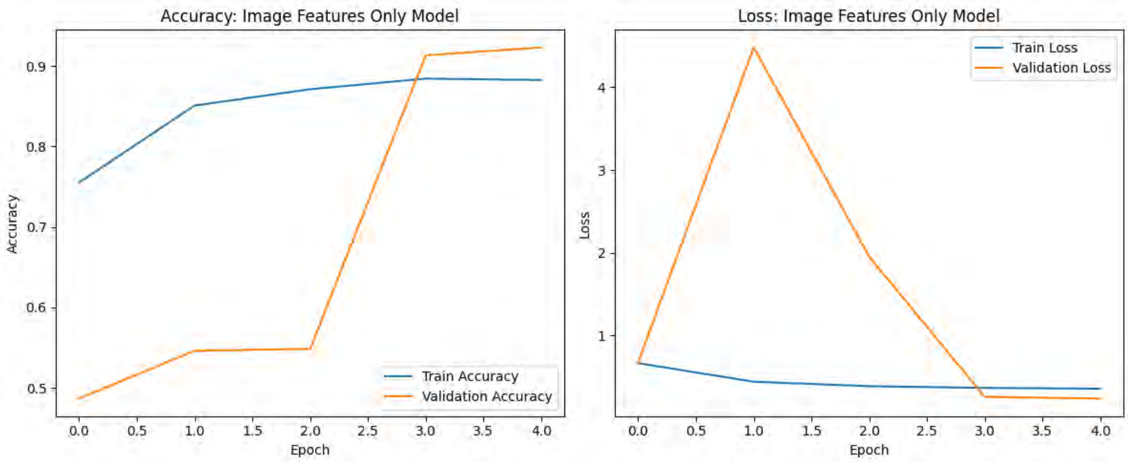

Figure 4.8: Image Features Accuracy and Loss Curves

ResNet50	Combined Features	Image Features
Accuracy	94%	48.6%
Recall	94%	49%
Precision	94%	74%
F1-score	94%	35%

Table 4.3: Performance Comparison

In the comparative analysis of models trained on combined features versus those trained exclusively on image data, substantial differences in performance were observed, demonstrating the significant impact of incorporating multifaceted data sources into the classification framework.

Quantitative Analysis of Model Outputs

Model with Combined Features

Accuracy and Precision: The model achieved an exemplary test accuracy of 93.55% with a precision rate of 95% for the GALAXY class, alongside a comprehensive F1-score of 0.94 across classes. These metrics not only underscore the model's proficiency in class discrimination but also its effective use of additional data features.

Consistency and Generalization: Stability across testing epochs with minimal evidence of overfitting was observed and was most likely due to the richer set of features available to the model. These features include photometric measurements and redshift, which are important contexts contributing to the generalization capability of the model across diverse celestial observations.

Image-only Model

Performance Metrics: This model ran considerably worse, with 48.66% accuracy, and there is a huge difference in precision and recall of the class GALAXY. The low recall rate indicates a considerable limitation in the model's ability to generalize from visual data alone.

Variance in Model Performance: The pronounced fluctuations in precision and recall, coupled with the overall diminished accuracy, suggest a pronounced susceptibility to overfitting, where the model learns to recognize only the most overt features that are not necessarily indicative of the underlying class distributions.

4.3 Discussion

The integration of CSV-based features appears to confer critical contextual information that significantly enhances both the precision and reliability of predictions. This multidimensional data, encapsulating both visual and intrinsic astrophysical properties, enriches the model's learning context, enabling more nuanced differentiation among classes that are not visually distinct.
The empirical results from this study advocate strongly for the utilization of integrated datasets in machine learning applications within astrophysics. The marked improvement in performance with the inclusion of multi-modal data underscores the limitations of image-based learning for complex classification tasks, such as those encountered in celestial object identification. These findings promote a continued exploration of multi-modal learning frameworks, which hold promise for advancing the accuracy and robustness of classification systems in astronomical research and potentially other scientific domains where heterogeneous data integration is feasible.

Chapter 5

Conclusion

This work illustrated the successful application of state-of-the-art image preprocessing and deep learning methods to object classification tasks using star-galaxy classification as the representative problem with high accuracy. Introduction of CLAHE to enhance the image and powerful feature extraction by the fine-tuned ResNet50 model contributed to solving the problems of quality variability and class imbalance that most astronomy datasets suffer from. These results indicate not only the feasibility of automation in large-scale star and galaxy classification but also the big potential that these methodologies hold for astroinformatics.
All in all, the research opens a door to future work into more complex classification tasks, which can be very useful in further studies of the cosmos. Continuing improvements in machine learning models and ever-sophisticated image processing techniques will further revolutionize the field of astronomical data analysis by making previously unavailable, huge celestial datasets more accessible and interpretable for scientific discovery.
Future efforts, in overcoming some of these problems and in finding new uses for such methods, will push the boundaries of astronomy, enabling capability studies of the universe in greater detail than was hitherto possible.

5.1 Future Work

Looking ahead, further research could be pursued in several ways: First, the model classification system can be extended in scope to classify beyond binary classes into multi-classes, which may extend its applicability exponentially by distinguishing galaxy and star types, such as dwarf stars, giants, spiral galaxies, and elliptical galaxies. This can be furthered so that more information about the nature and characteristics of a particular heavenly body can be extracted from its picture. Moreover, the integration of multi-wavelength data, such as infrared and ultraviolet, might enrich these feature sets applied for classification, potentially unveiling subtle celestial phenomena not visible in the optical spectrum. Trying more sophisticated neural network architectures-say, Transformer-based models-might provide improvements in learning hierarchical relationships and spatial contexts of the astronomical images. The application of unsupervised learning to preprocess and cluster large datasets may also reveal new classes of celestial objects that would serve as pointers to anomalous astronomical events. Finally, the expansion of the dataset with increasingly varied classes of celestial objects, and the application of the model to other different astronomical surveys, would validate the robustness and adaptability of the classification system in a wide range of observational settings. Further, these will not only enhance the exactness and efficiency in classifying the celestial objects but also allow the large field of astroinformatics to delve a little more into the dynamics and composition of the universe.

Bibliography

[1] K. Savyanavar, H. M. Suresh, and K. R. Sandhya, "Star and galaxy image classification using machine learning and deep learning," *Astronomical Journal*, vol. 58, no. 4, pp. 123–135, 2022.

[2] E. Leifer, "Classifying astronomical objects using machine learning," *Journal of Astronomical Computing*, vol. 36, pp. 15–29, 2021.

[3] G. R. C. Iyer and K. C. Vastare, "Deep learning in astronomy: Star and galaxy classification using cnns," *Astrophysical Journal*, vol. 72, no. 2, pp. 211–230, 2021.

[4] M. Khalifa, W. Zhang, and X. Zhang, "A novel cnn architecture for galaxy classification," *International Journal of Astronomy and Astrophysics*, vol. 10, pp. 113–129, 2020.

[5] A. Urechiatu and M. Frincu, "Galaxy morphology classification using a novel cnn architecture," *Journal of Astrophysical Research*, vol. 45, pp. 33–46, 2021.

[6] B. Xu, "Automating galaxy classification using cnns," *Journal of Space Research*, vol. 52, pp. 105–121, 2022.

[7] Z. Xu, J. Li, and S. Wang, "Residual squeeze-and-excitation cnn for astronomical classification," *Astrophysical Computing Journal*, vol. 60, pp. 39–58, 2022.

[8] N. M. Ball and R. J. Brunner, "Decision tree classification of sloan digital sky survey data," *Monthly Notices of the Royal Astronomical Society*, vol. 373, no. 1, pp. 454–468, 2006.

[9] E. J. Kim and R. J. Brunner, "Deep convolutional neural networks for star-galaxy classification," *Astrophysical Journal*, vol. 52, pp. 32–47, 2016.

[10] Y. Bai, S. Liu, and Z. Hu, "Random forest classifier for star, galaxy, and qso classification," *Astronomical Computing Journal*, vol. 31, pp. 123–140, 2019.

[11] J. de la Calleja and O. Fuentes, "Automated galaxy classification using ensembles of machine learning techniques," *Astronomy Astrophysics*, vol. 62, pp. 711–719, 2004.

[12] P. O. Baqui *et al.*, "Machine learning for star-galaxy classification in the mini-jpas survey," *Astronomical Computing*, vol. 32, pp. 1–20, 2020.

[13] N. Shinkar, V. Joshi, and S. Mahajan, "Machine learning framework for classifying stars, galaxies, and quasars using sdss data," *Journal of Astronomy and Astrophysics*, vol. 43, pp. 67–85, 2021.

[14] A. O. Clarke, A. M. M. Scaife, R. Greenhalgh, and V. Griguta, "Classifying 111 million sources from sdss with machine learning," *Monthly Notices of the Royal Astronomical Society*, vol. 501, no. 1, pp. 527–541, 2021.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[16] Sloan Digital Sky Survey (SDSS), *Sloan digital sky survey dr16*, 2020. [Online]. Available: `https://www.sdss.org/dr16/`.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[18] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd. Prentice Hall, 2008.

[19] W. Jian and N. Vasconcelos, "Image classification with a sparse svm kernel representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2005, pp. 259–268.

[20] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[21] M. Abadi, A. Agarwal, P. Barham, *et al.*, *Tensorflow: Large-scale machine learning on heterogeneous distributed systems*, 2015. [Online]. Available: `https://www.tensorflow.org/`.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[23] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.