Astronomical Image Classification using Machine Learning for Point Source Image Detection

Bachelor of Technology

with specialization in

Information Technology

Submitted by

Parth Revanwar	IIB2022044
Nilesh Chaubey	IIT2022261
Vraj Shah	IIT2022263
Shashank Arora	IIT2022502
Mohit Bajaj	IFI2022015

Under the Supervision of

Prof. Pavan Chakraborty

to the

Department of Information Technology

Indian Institute of Information Technology,
Allahabad

भारतीय सूचना प्रौद्योगिकी संस्थान इलाहाबाद

Indian Institute of Information Technology Allahabad

An Institute of National Importance by Act of Parliament

Deoghat Jhalwa, Prayagraj 211015 (U.P.) India
Ph: 0532-2922025, 2922067; Web: www.iiita.ac.in; Email: contact@iiita.ac.in

CERTIFICATE

It is certified that the work contained in the thesis titled "Astronomical Image Classification using Machine Learning for Point Source Image Detection" has been carried out under my/our supervision and that this work has not been submitted elsewhere for a degree.

Prof. Pavan Chakraborty
Department of Information Technology
IIIT Allahabad

भारतीय सूचना प्रौद्योगिकी संस्थान इलाहाबाद
Indian Institute of Information Technology Allahabad
An Institute of National Importance by Act of Parliament
Deoghat Jhalwa, Prayagraj 211015 (U.P.) India
Ph: 0532-2922025, 2922067; Web: www.iiita.ac.in; Email: contact@iiita.ac.in

CANDIDATE DECLARATION

I, Shashank Arora, Roll No. IIT2022502, certify that this thesis work titled "Astronomical Image Classification using Machine Learning for Point Source Image Detection" is submitted by me towards partial fulfillment of the requirement of the Degree of Bachelor of Technology in the Department of Information Technology, Indian Institute of Information Technology, Allahabad.

I understand that plagiarism includes:

1. Reproducing someone else's work (fully or partially) or ideas and claiming it as one's own.
2. Reproducing someone else's work (verbatim copying or paraphrasing) without crediting.
3. Committing literary theft (copying some unique literary construct).

I have given due credit to the original authors/sources through proper citation for all the words, ideas, diagrams, graphics, computer programs, experiments, results, and websites that are not my original contributions. I have used quotation marks to identify verbatim sentences and given due credit to the original authors/sources.

I affirm that no portion of my work is plagiarized. In the event of a complaint of plagiarism, I shall be fully responsible. understand that my supervisor may not be in a position to verify that this work is not plagiarized.

______________________________ Date: ______________

Shashank Arora
IIT2022502
Department of Information Technology
IIIT Allahabad
Prayagraj - 211015, U.P.

"Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I also got a good job in industry."

Shashank Arora

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD

Astronomical Image Classification using Machine Learning for Point Source Image Detection

ABSTRACT

Bachelor of Technology
Department of Information Technology

by Shashank Arora

his research focuses on the classification of astronomical images using point source image detection to distinguish between stars, galaxies, and other celestial objects. Point sources, which appear as concentrated light sources in telescopic data, present a challenge due to their similar visual structures. We explore various machine learning models, including logistic regression, random forests, and convolutional neural networks (CNNs), to address this challenge. By comparing their performance on a curated dataset, we demonstrate the effectiveness of CNNs in accurately classifying point sources, offering significant improvements in the accuracy of astronomical object detection and classification.

Acknowledgements

We would like to express our heartfelt gratitude to **Prof. Pavan Chakraborty** for his invaluable guidance, constant encouragement, and insightful advice throughout the course of our mini project titled "Astronomical Image Classification using Machine Learning for Point Source Image Detection." His expertise and suggestions were pivotal in helping us navigate the challenges we encountered, and we deeply appreciate his support.

We would also like to extend our sincere thanks to **Dr. Snigdha Sen**, who served as our guide during the project. Dr. Sens guidance played a crucial role in keeping us on the right path whenever we faced difficulties. Her timely advice and direction were instrumental in helping us understand complex concepts and tackle problems effectively. We are grateful for her patience and mentorship throughout this journey.

As 5th semester B.Tech students, we found the process both challenging and rewarding. The project allowed us to enhance our knowledge in machine learning and its applications in astronomy, a field we had not previously explored. We thoroughly enjoyed working on this project and appreciate the learning experience it provided.

Once again, we extend our sincere thanks to Prof. Pavan Chakraborty and Dr. Snigdha Sen for making this project a success.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

ABSTRACT

In this research, we focus on the classification of astronomical images using point source image detection, a method that isolates small, concentrated light sources such as stars and distant galaxies. The challenge of distinguishing between different types of point sources, especially between stars and galaxies, presents a unique problem in astronomical image analysis due to their visual and structural similarities.

This chapter provides an introduction to the research problem, beginning with an overview of the significance of astronomical image classification in modern astronomy. We then explore the concept of point source images, which represent objects appearing as compact light sources in a telescopes field of view. Understanding the characteristics of these point sources is critical for accurate classification. This introduction lays the foundation for the methods and models used in this research to improve the accuracy of point source classification.

1.1 Introduction

Sir William Herschel, the famous astronomer and composer, once said, "The method I have taken of analyzing the heavens, if I may so express myself, is perhaps the only one by which we can arrive at a knowledge of their construction " [11] [Her89]. Herschel then tried to catalog the heavenly bodies which he observed in the night sky , providing a short description of each nebula or cluster of stars, as well as its situation with respect to some known object hoping that this would one day encourage other like minded astronomers and induce them to undertake the necessary observations. Generations since have indeed tried to explore the vast unknowns of the universe and have made significant leaps and bounds in recognising and categorizing different types of stars, galaxies and quasars. A quasar is an extremely luminous active galactic nucleus (AGN) [24].

However this race to classify and learn about these new discoveries and advancement in technologies have led to one peculiar problem: We have more data than we can handle.

The launching of new telescopes with increasing resolutions have bombarded astronomers with humongous amounts of data on enormous numbers of astronomical objects, orders of magnitude more than they can catalog through their own observations and analyses. Moreover, classification of images of stars and galaxies based on morphology alone according to which unresolved point sources are classified as stars, and resolved sources for which a shape can be determined are classified as galaxies has led to inaccurate classifications [3][Cea20a]. Using this approach, images of quasars, galaxies with an extremely luminous center region, particularly if they are very distant, are sometimes mistaken for stars [3].

In this paper , we have used different methods of machine learning from very basic to fairly advanced, in hopes of striking a balance between complexity of model and its accuracy. Due to our limited experience and knowledge of Machine Learning and Deep learning, these models may seem very rudimentary at face value. However we have tried our best to learn various nuances and techniques pertaining to this domain. The algorithms used can be used to classify images of stars and galaxies without prior feature extraction. To keep things simple and efficient, we have trained our models using 64x64 pixelated images of stars and galaxies without inputs of additional measurements of the brightness, size, or shape of stars or galaxies.

We have implemented 3 different classification algorithms of increasing complexities: A Logistic regression model as the baseline benchmark, Random Forest Classifier as an advanced ML model, and have finally ventured into the unknown depths(for us) of Convoluted Neural Networks (CNN) to try and beat conventional Machine learning models.

1.2 Point Source Images

Point source images refer to those images which are characterized by their appearance of a small concentrated point of light source or reflection. In physics or optics, a point source is a mirage of a very large object located far away from the lens and which can be considered as an ideal object whose size is negligible when compared to the systems resolution. These sources emit light or other radiation uniformly in all directions.

Stars, galaxies and other bodies often appear as point sources from Earth in the sense that these objects are mind boggling distances away from us , and thus appear as insignificant noise dots in images, but in actuality are of utmost importance. They play a crucial role in astronomy as detecting and classifying them is a generally tedious task due to the limited amount of information available in their photographs.

1.3 Stars and Galaxies

A star is a massive ball of hot gas [1][10] . Typically, stars consist primarily of hydrogen and helium gas, with small amounts of other elements many stars are about 73 percent hydrogen gas, 25 percent helium gas, and 2 percent other elements [8] . The lifetimes of stars is incomprehensible in the sense that an average star has a lifespan ranging from a few million years to billions of years. Stars are classified based on their various features such as mass, temperature, and brightness.The ages and compositions of stars in a galaxy can provide information about the history, dynamics, age, and evolution of the galaxy .

A galaxy is a massive cluster of stars, gaseous clouds, and dust, all held together by gravity, without which chaos will ruin the universe. Astronomers classify galaxies into three major categories: elliptical, spiral and irregular. These galaxies span a wide range of sizes, from dwarf galaxies containing as few as 100 million stars to giant galaxies with more than a trillion stars. [13]

The problem with classifying galaxies and stars is that photos of some galaxies closely resemble stars, as they have very bright center spots which overshadow their outer structure, thus masquerading them as stars. The extremely bright centers of these galaxies, with a brightness that cannot be accounted for by the presence of stars alone, are called active galactic nuclei or AGNs [12] [21]. These AGNs are often referred to as Quasars. These Quasars are many a times brighter than galaxies. For example, a particular quasar is a thousand times brighter than our own galaxy, The Milky Way. These make them some of the brightest structures in the universe and consequently our images . As a result, distant quasars can be mistaken for stars, even using modern machine learning techniques [23][4] .

Chapter 2

Literature Review

ABSTRACT

This chapter provides a thorough survey of the problems identified in the field of Classification of Astronomical images . In addition, a number of challenges were identified in the previous chapter which forms the basic motivation of research towards solving this problem using various algorithms. Multiple research papers were thoroughly studied by us and a detailed analysis about the methodologies used, datasets , drawbacks, advantages and relevance to our research topic was carried out for each paper. The review of these literature provided us with many valuable insights into solving this problem and helped us formulate a plan of action towards solving our given problem statement.

2.1 Paper 1 - Astronomical Images Classification Using Deep Learning CNNs [18]

Introduction

In this research paper author explores the use of cnn for classification of astronomical images as supernovae, quasars, and artefacts or more broadly real and not real the vast dataset used in the training is gathered from various sky surveys and hard to study manually.

Methodology

The authors used cnn for classifying the object as real or non real such as noise or artefact the model used python with keras and used 3 convolutional layers with increasing filter size succeed by Max Pooling and ReLU activation functions. Sigmoid function is used in final layer to classify data in binary format the loss function used in binary classification is cross entropy and training process done using RMSprop optimizer

Dataset

The Authors used data from Sloan Digital Sky Survey (SDSS) ,Supernova Survey,SDSS-II SN Survey which was conducted between 2005-08 . the dataset contains 20k+ images in 69*69 size in pixel .The dataset used in this model is available to use publicly .the data contains 10 classes but were combined to only 3 such as real artefacts ,dipole

Drawbacks

Limited classification(only binary) : Only partial classification the model is only trains to classify data in binary as real or noise objects its do not help in classifying the real objects in different classes Overfitting : The second major issue with the model was overfitting: the authors stated that model was overfitting on test data and was less accurate on unseen data

Advantages

Data preprocessing for multi classification :Although the model is not helpful to classify in multiclasses it is very useful to classify data for future model to be more precise on multi classification .

Relevance of this paper

Our main goal is to classify point source objects in different classes rather than only detecting noise and real objects so we can divide our dataset from this model to remove noise and make data set relevant to our goal.

2.2 Paper 2 - "Principled Point-Source Detection in Collections of Astronomical Images" [16]

Methodology

This paper aimed to overcome some of the common difficulties faced by scientists during processing and analysis of astronomical images like an isolated source in background-dominated imaging with perfectly known background level, point-spread function, and noise models. The authors of this paper have identified the matched filter method for the detection of point source in images and their classification into various celestial bodies like quasars, stars, galaxies and many more. This method has been previously worked upon in many papers.

Test Data

The paper uses the Supernova program of the Dark Energy Survey (DES) from the Dark Energy Camera (DECam) dataset. The above includes images from various bands like the g, r and i ,with varying exposure times and noise levels. The images were then calibrated with the NOAO DECam Community Pipeline for preprocessing.

Public Availability: The above dataset used in the paper is publicly available to all , which allows further work and contribution to this project.

Results

The authors have used the already existing matched filter technique and shown how it can be harnessed to detect and classify point sources in astronomical images. This method was further extended to SED space in order to combine images taken through multiple bandpass filters.

Relevance

Relevance to our paper : This paper shows significant relevance to our research topic and provides essential information about the topic. It also helped us significantly in learning new techniques and provided new insights into the topic. While this paper has utilized a matched filter method, our approach harnesses machine learning and deep learning models, which could potentially offer a more flexible solution by learning directly from data.

Drawbacks

One of the drawbacks of this matched filter method is that it is highly dependent on ideal conditions such as known PSFs and noise models. As we know, in real life these ideal conditions are very hard to match, and hence this method could suffer in performance . On the contrary , Deep learning approaches work on more general data adapting to more complex and noisy data without hard and fast assumptions and constraints

Advantages

Mathematical Optimality: The matched filter provides the best possible detection given the assumptions. For applications requiring high precision, this method might outperform machine learning models, especially in well-understood datasets.

2.3 Paper 3 - Deep Learning Techniques for Astronomical Object Classification [6]

Methodology

This paper, titled "Deep Learning Techniques for Astronomical Objects Classification" utilizes a structured approach in the form of the KDD (Knowledge Discovery in Databases) methodology for the classifying and analyzing point sources in astronomical images. It explores and uses several deep learning models like VGG16, InceptionV3, ResNet50. Pre-trained weights were then applied to these models through transfer learning from ImageNet to classify objects such as galaxies, stars, and quasars.The model was then fine tuned .

Test Data

The test data of this research was taken from Sloan Digital Sky Survey (SDSS) Data Release 17 (DR17). The data, which was downloaded in FITS (Flexible Image Transport System) format , consisted of various images covering many classes of celestial objects, which was then converted prior to processing into JPEG format.

Results

The main model used in this paper, which was the VGG16 model, achieved an impressive accuracy of 86.04% and proved to perform the best. This was followed by the InceptionV3 model which provided an accuracy of 83.92%.Other models like the ResNet50 and CNN models scored 79.79% and 79.57%, respectively.

Advantages

This paper is very related to our research topic, as it tackles the same issue: how to classify astronomical objects with so-called deep learning techniques of the type of CNN architectures.

Drawbacks

The study is almost entirely concerned with galaxies, stars and quasars; other relevant astronomical objects like supernovae or nebulae were not taken into consideration for classifying. Moreover, the number of epochs is 10, which could be insufficient in order to complete the learning process by the model.

2.4 Paper 4 - Detection and Classification of Astronomical Targets with Deep Neural Networks in Wide-field Small Aperture Telescopes [15]

Methodology

The authors have studied the astronomical target detection and classification framework which utilizes deep neural networks (DNNs),which focuses on data from wide-field small aperture telescopes (WFSATs).The main algorithm used is R-CNN with a modified Resnet-50 and Feature Pyramid Networks (FPN) for detecting and classify celestial objects in images faster.

Test Data

The testdata that has been used includes both simulated data (generated using the SkyMaker tool) and real observational data from WFSATs. We generate simulated data to match the conditions of WFSATs, which helps in tracking celestial objects in sidereal mode. The prime focus of the paper to detect point-like sources such as stars, supernovae, comets, and meteors.

Results

This model shows better results for the detection of faint sources as compared to traditional models. It is capable of detecting dimmer objects more efficiently with higher precision in handling streak-like objects compared to classical detection algorithms. Using both simulated and real data allows the framework in generalizing better and hence achieving 94% classification accuracy for 8 various celestial objects.

Relevance and Drawbacks

Relevance: The papers concentrate on applying machine learning to classify astronomical targets. Our research focuses on point source image recognition, whereas their paper deals with streak-like and point sources; so, while they have similar goals, they differ in particular target kinds. **Cons**: This paper's emphasis on multi-type object categorization rather than just point-source identification is a possible negative in comparison to our focus, which may result in less optimization.

2.5 Paper 5 - Image Classification of Stars and Galaxies Using Different Machine Learning Models" [17]

Methodology

We are using the differences in Spectral Energy Distributions(SEDs) of the images of stars and galaxies where they seem to be morphologically similar. Our models are trained by using 3986 images (64x64 pixel cutouts) using one-hot encoding and scaling the pixel values. We used hyperparameters such as nodes,layers and kernel size for tuning the CNN model, and for the Random Forest model, trees and minimum samples per leaf were adjusted. We trained the neural network model using a dataset preprocessed with PCA.

Dataset

The images that our dataset contain were captured by the 1.3m Devasthal Fast Optical Telescope (DFOT) with a 2048 x 2048 pixel grid. It was reduced to 64 x 64 cutouts to show a single astronomical object. Our dataset contains 3986 images: 3044 images of stars, and 942 images of galaxies.

Results

The Neural Network with PCA performed the best, achieving an accuracy of 84%. We analyzed that CNN model with kernel size of 6 performs the best with an accuracy of 82.6 percent. Logistic Regression and Random Forest models showed underwhelming accuracy of 73% and 79% respectively.

Relevance to Our Paper

We found that using machine learning models can be effectively used to classify images of stars and galaxies, saving the time required to catalog them. The CNN model uses a pooling layer which is placed after convolution that reduces the number of parameters thus reducing the overall models runtime.

Drawbacks

Limited Model Scope: The neural network approach with PCA may face some problems while dealing with more complex or higher-resolution data. Moreover, we observe the overfitting issue in CNN that highlights a need for a more robust data augmentation or other architectures.

Advantages

Diverse Methods Tested: By studying multiple machine learning models, we get to know about a variety of strategies for astronomical image classification.

The use of PCA is a helpful technique for reducing complexity, which could improve model generalizability.

2.6 Paper 6 - "DeepSource: Point Source Detection using Deep Learning " [[22]

Methodology

In the following paper the author has presented DEEPSOURCE a deep learning solution that uses convolutional neural networks to enhance the signal to noise ratio of the original image. This technique has outperformed the traditional Machine learning based approach which was using feature extraction and prediction analysis . A Neural Network with multiple layers is utilized which gives us the advantage of automatically learning representative features from the data directly.

Test Data and Model

The model was trained on two sets of classes - Same-Field and Different-Field, each with 500 images of size 1 Œ 1. The dataset was taken from STIMELA MeerKat images. Natural Weighting and Uniform Weighting are used to correct the sky brightness distribution and large u-v gaps. It is used to vary the robustness, R which is approximated to be 1.5. Four highly powered GPUs were used to perform 10K iterations for the deconvolution of each image.

Results

Purity and Completeness were kept in mind while preparing the results for the following model. It produces higher purity and completeness scores, with a PC (Purity Œ Completeness) score of 0.73 at SNR = 3, compared to PYBDSFs 0.31. This indicates that DeepSource detects fainter sources more effectively and with fewer false positives.

Relevance and Drawbacks

Both DeepSource and our paper deal with point source detection in astronomical images using machine learning techniques. DeepSource uses CNNs for noise suppression and source detection, which is in alignment with our goal of classifying point sources. The use of deep learning for SNR enhancement is a strength of DeepSource.The models reliance on large, complex CNN architectures may not be ideal for real-time applications or resource-constrained environments.

Chapter 3

Present Investigation

ABSTRACT

This chapter presents the methodology for classifying astronomical point source images using various machine learning models. The Dataset consists of labeled images of stars and galaxies. We begin with Logistic Regression as a baseline model, followed by the Random Forest classifier, which improves accuracy through an ensemble of decision trees. Finally, we explore Convolutional Neural Networks (CNNs), which automatically extract features from raw images, offering superior performance. The chapter compares these models, demonstrating CNNs effectiveness in accurately classifying celestial objects based on their point source characteristics.

3.1 Dataset

The dataset used by us in this paper was obtained from Kaggle- where it is publicly available for use and download.It is named - "Star-Galaxy Classification Data" [7]. This data was created as a part of a project at Aryabhatta Research Institute of Observational Sciences (ARIES), Nainital, India. The images were captured by the in-house 1.3m telescope of the observatory situated in Devasthal, Nainital, India. The original images captured were 2kx2k in size which was reduced to 64x64 cutouts from the images to isolate the sources in a single image [7] . For labelling the images, the images were first passed through image segmentation to identify various sources in the images and finally the center coordinates of the found sources were queried with the SDSS database to give a label corresponding to each 64x64 cutout. This is a binary classification dataset containing images of Stars and Galaxies.

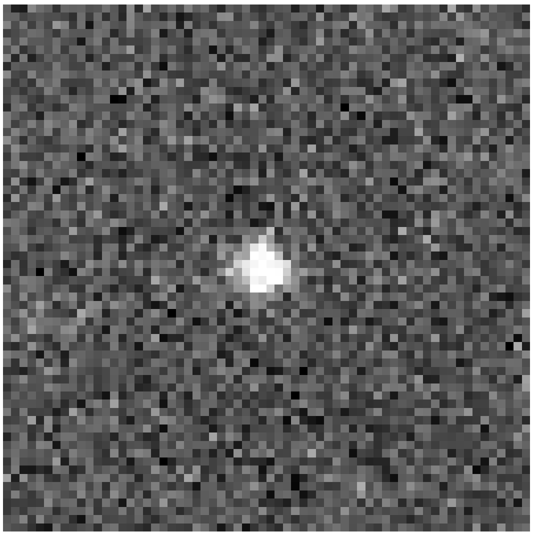

FIGURE 3.1: Star

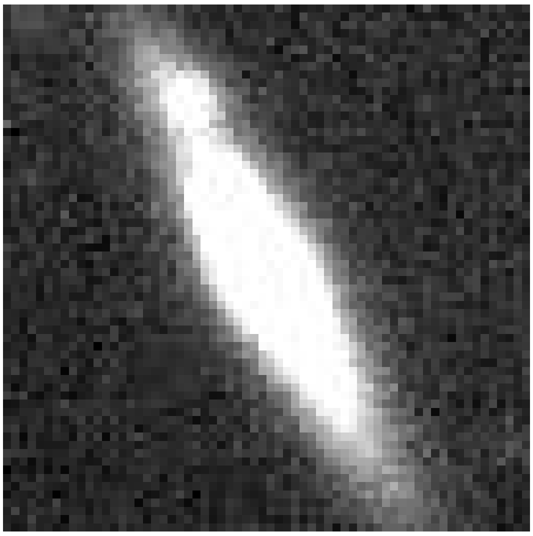

FIGURE 3.2: Galaxy

3.2 Data Preprocessing

Image preprocessing is the process of manipulating raw image data into a usable and meaningful format. It allows you to eliminate unwanted distortions and enhance specific qualities essential for computer vision applications [19]. Data preprocessing is an very important step in machine learning ,specifically when working with image data in machine learning images are represented as multidimensional vector of pixel ,where each pixel is important for presentation of image .so various preprocessing steps are needed to ensure data is suitable for analysis steps include loading ,resizing in fixed dimension for consistent dataset ,converting it into a 1 D vector to simplify input ,giving labels for classification . also balancing the data for each class and ensure shuffling for appropriate results

1. **Image resizing** - Each image is resized to 128*129 pixel .this is to ensure each input has same size and easy for ML model to understand.

2. **Flattening the image** - Each image is the converted to a multidimensional vector and the flattened into 1 single dimensional for easy access for machine learning model.
3. **Balancing the data set** - It is crucial to ensure that the no of classes in which we ar edividing the data should contai nearly equal no of images or the model is inclined to some specific classification due to lareg number of images of one particular type of data

3.3 Logistic Regression

Logistic regression estimates the probability of an event occurring, such as voted or didnt vote, based on a given data set of independent variables.[14] It is a binary classification model, where the aim is to classify the data points into two different categories. (e.g., spam vs. not spam, cancerous vs. non-cancerous). Logistic regression is similar to the neural network without hidden layers, meaning it provides a mechanism that directly maps the inputs to outputs avoiding any complex transformations.

It is simple as it has lesser parameters .Example, in this case 4096 parameters are used.Also it does not contain hyperparameters such as learning rate, momentum, or activation functions which are present in complex models like Convolutional Neural Networks (CNNs).

3.3.1 Binary Cross-Entropy Loss Function

The binary cross-entropy loss function is utilized by logistic regression because it is a binary classifier. The difference between the true binary labels and the expected probability is measured by this loss function.

Since binary cross-entropy penalizes inaccurate predictions based on the confidence in those forecasts, it is frequently utilized in models where the output is a probability between 0 and 1.

3.3.2 Overfitting in Logistic Regression

Despite the simplicity of logistic regression, the model was overfitted to the training data. Even though logistic regression is a simple model, still the model gets overfitted to the training data. When a model works incredibly well on training data but is unable to generalize to fresh, untried data, this is known as overfitting. In this instance, the logistic regression model's accuracy on the training set was 100%, while it was only 73% on the test set. This is surprising for such a basic model because it implies that the machine memorized the training data instead of learning the generalizable patterns. To avoid this we can use regularization.

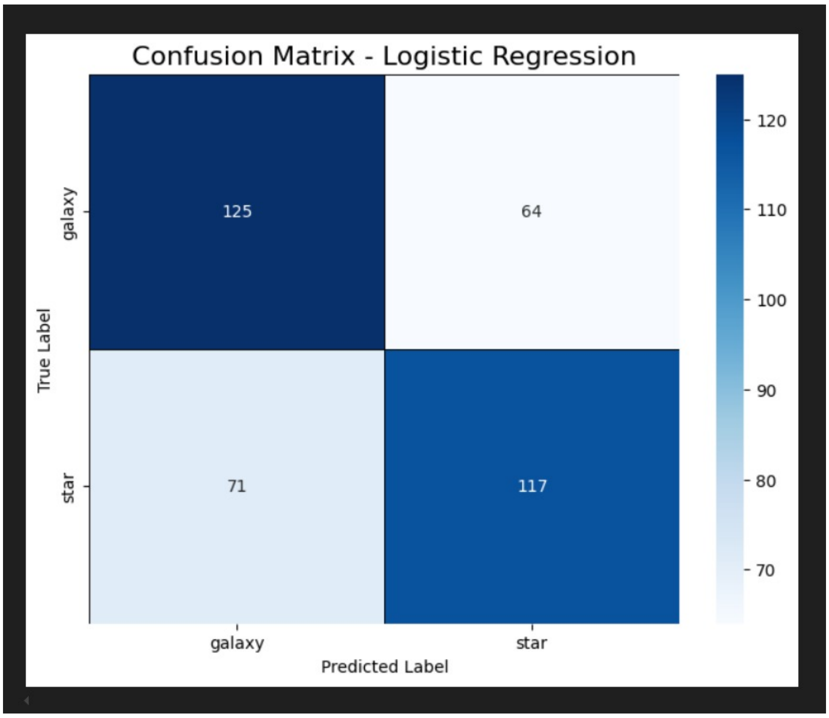

FIGURE 3.3: Confusion Matrix

3.3.3 Number of Parameters

In this example we have used less number of parameters. However, overfitting still happened even with this modest number of parameters, most likely as a result of the dataset's complexity. Despite being simpler, logistic regression can still overfit, especially in cases where the training data is relatively small or complex. While the model has limitations, logistic regression remains a useful tool in cases where interpretability, simplicity, and efficiency are more important than raw performance, especially in situations where complex models like CNNs are overkill or computationally expensive. Hyperparameter tuning is an optimization technique and is an essential aspect of the machine learning process. A good choice of hyperparameters may make your model meet your desired metric [5] .

3.3.4 Results and Conclusion

TABLE 3.1: Classification report for Logistic Regression Model

	precision	**recall**	**f1-score**	**support**
0	0.64	0.66	0.65	189
1	0.65	0.62	0.63	188
accuracy			0.64	377
macro avg	0.64	0.64	0.64	377
weighted avg	0.64	0.64	0.64	377

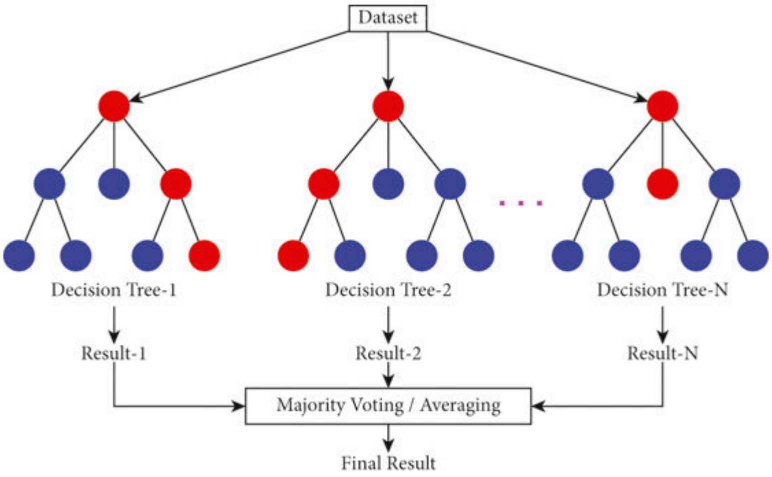

FIGURE 3.4: Random Forest

3.4 Random Forest

It is an ensemble learning model which uses multiple decision tress to increase accuracy of decision trees .It uses the idea that a group of weak learners come together to crete a string learner .Random forest is a supervised learning algorithm. The forest it builds is an ensemble of decision trees, usually trained with the bagging method [2] .

Random Forest uses a technique called bagging, where multiple subsets of the training data are created by sampling with replacement. This means some data points may be repeated in a subset while others may not be included. For each subset of data, a decision tree is constructed. During the construction of each tree, only a random subset of features is considered for splitting at each node. This randomness helps in reducing overfitting. For classification tasks, each tree in the forest votes for a class label, and the class with the majority of votes becomes the final prediction.

3.4.1 Hyperparameters

The two hyper parameters in the random forest are number of trees and minimum samples per leaf. The optimal value for balancing performance of model and efficiency to calculate is 100 trees. Minimum samples per leaf to 8 to reduce overfitting while also maintaining details in tree splits. Upon further testing of the code, we could generate the following accuracy by varying the hyperparameters.

3.4.2 Performance

The accuracy for random forest classification was calculated to be 73%.

n_estimators	max_depth	min_samples_leaf	mean_test_accuracy
100	20	10	0.753169
200	None	8	0.753168
100	30	5	0.752502
200	10	8	0.751182
50	10	8	0.750514
100	10	5	0.750508
200	20	8	0.748757
200	30	8	0.748753
200	30	5	0.747195
200	None	10	0.747195

TABLE 3.2: Model Accuracy by varying Hyperparameters

	precision	recall	f1-score	support
0	0.71	0.74	0.73	185
1	0.74	0.71	0.73	192
accuracy			0.73	377
macro avg	0.73	0.73	0.73	377
weighted avg	0.73	0.73	0.73	377

TABLE 3.3: Classification Report for Random Forest Model

3.4.3 Strengths

This model is robust to overfitting helping to generalize to unseen images. The classification uses high dimension of data which is effectively processed by random forest. Gives idea about which features are important to distinguish stars from galaxies.

3.4.4 Weakness

It takes longer training time due to number of trees which may be hard to train for big data. It is less interpretable compared to decision tree.

3.4.5 Gini Index

How nodes on the decision tree branch the given formula uses class and probability to determine gini of each branch on a node pi represents the relative frequency of the class you are observing in the dataset and c represents the number of classes.

$$Gini = 1 - \sum_{i=1}^{C} p_i^2$$

3.4.6 Entropy

We can also use entropy to calculate how nodes branch in a decision tree. Entropy uses probability of some outcome to find which node to branch.

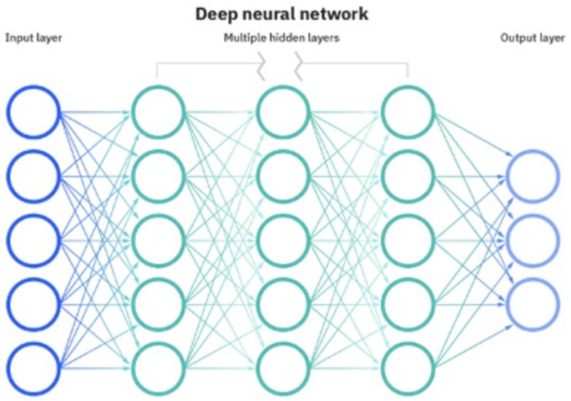

FIGURE 3.5: Neural Network

$$Entropy = \sum_{i=1}^{C} -p_i \log_2(p_i)$$

3.4.7 Conclusion

The random forest classifier performs strongly in classifying galaxy and stars with an accuracy of 79Robustness and complexity is balanced well in this model spatially in the case of astronomical images

3.5 Convolutional Neural Network

3.5.1 Neural Network

Neural Networks form the basis of Deep Learning. Neural Networks are computational models that mimic the complex functions of the human brain. The neural networks consist of interconnected nodes or neurons that process and learn from data, enabling tasks such as pattern recognition and decision making in machine learning.[9].These networks are ultimately just interconnected nodes, called neurons which are organized into layersinput, hidden, and output layers. Each neuron processes input data, applies some transformation, and transfers the information to the next layer. For instance, the initial layers can start with identification of edges and then the further layers can detect the complex parameters which are in turn a combination of edges and corners of stars and galaxies. The number of layers and nodes in each layer can all be considered as hyperparameters. Once these hyperparameters are determined each node can be represented as a combination of features present in it.

3.5.2 CNN

A Convolutional Neural Network (CNN) is a type of deep learning neural network model specifically designed for processing visual data which is represented in the form of matrices. There are three primary operations of CNNconvolution, pooling, and fully connected layers to learn patterns in data. Unlike traditional neural networks, which treat all pixels equally, CNNs preserve the spatial structure of images ensuring that each pixels intensity and its respective position both are kept in consideration, making them exceptionally powerful for image classification tasks. The following tasks are performed by these three layers -

1. Convolutional layers apply filters across the input image to extract features like edges, textures, or more complex structures.
2. Pooling layers downsample the feature maps, reducing dimensionality while retaining critical information.
3. Fully connected layers perform the final classification based on the features extracted by earlier layers.

3.5.3 Why Deep Learning over Machine Learning

Classical machine learning models such as Support Vector Machines (SVMs), k-Nearest Neighbors ,Decision Trees and Logistic regression require fixed predetermined features to represent the input data, which can be challenging and time-consuming, especially for complex tasks like astronomical image classification. In Machine Learning the model depends on features such as shapes, textures, or colors manually. In contrast, CNNs automate this process by learning relevant features directly from raw images given to it. Deep Learning is gaining much popularity due to its supremacy in terms of accuracy when trained with huge amount of data [20]. Through the use of multiple layers, CNNs capture both low-level features like edges and high-level features like object shapes, thus eliminating the need for manual feature engineering. Another reason for shifting our focus towards CNN is that astronomical images are often large and contain high-dimensional data. Classical ML models struggle with high-dimensional data because they are sensitive to irrelevant features and may suffer from the "curse of dimensionality," which leads to performance degradation and overfitting on the test data. CNNs

Chapter 4

Timeline and Deliverables

ABSTRACT

This chapter presents the further plan and the timeline that we would follow to complete the project in due time. We are planning to implement a novel algorithm.

4.1 Further Plan

4.1.1 Initial Research work and collecting sample data

We searched for research papers to gain some idea about the topic , Astronomical Image Classification using Machine Learning for Point Source Image Detection and collected relevant information from various surveys, to understand the problem space and identify suitable datasets for model training and evaluation.

4.1.2 Implementing Logistic Regression to train the model

We implemented our model using logistic regression as our first approach because of its simplicity and efficiency. We got a rough idea about our test data and the overall workflow. It also helped us to test the data quickly without the need of complex resources. But due to inefficient performance with high-dimensional image data , not able to capture complex relationships and limited performance with non-linearly separable -data. Hence ,to overcome the limitations of logistic regression, we switched to Random Forest, an ensemble-based machine learning model.

4.1.3 Transition to Random Forest Algorithm

To overcome the limitations of logistic regression, we switched to Random Forest, an ensemble-based machine learning model. It was able to handle the non linear relationships among the data efficiently.The accuracy and the robust nature were improved due the comination of multiple descision trees. We were also able to identify significant image features because of feature importance analysis. But the model struggled to directly process raw image data without pre-extracted features.

4.1.4 Planning to implement Novel Algorithm using Deep Learning

To overcome the limitations of the Random Forest, we are planning to implement a novel algorithm using CNN algorithm to our classification model. At present we have limited knowledge about the algorithm, So we are planning to spend one-two weeks of time to gain sufficient knowledge about the algorithm. Further, we will be practically implementing the classification algorithm on the dataset as we did for the traditional classifiers like Logistic Regression and random forest and will be studying the thoroughly analysing the accuracy and computational cost. Further , we will implement a novel model for the classification of astronomical images and will be presenting it during the end -semester evaluation.

Bibliography

[1] International Astronomical Union Office of Astronomy for Education. *Glossary term: Star*. https://astro4edu.org/resources/glossary/term/331/. Accessed: 22/09/2024.

[2] BuiltIn. *Random Forest Algorithm*. https://builtin.com/data-science/random-forest-algorithm. Accessed: 23/09/2024.

[3] A.O. Clarke et al. "Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra." In: *Astronomy and Astrophysics* 639 (2020), A84.

[4] A.O. Clarke et al. "Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra." In: *Astronomy and Astrophysics* 639 (2020), A84.

[5] Codex. *Do I Need to Tune Logistic Regression Hyperparameters?* https://medium.com/codex/do-i-need-to-tune-logistic-regression-hyperparameters-1cb2b81fca69. Accessed: 2024.

[6] Yogiraj Subhash Dalvi. "Deep Learning Techniques for Astronomical Object Classification." Submitted. MA thesis. Dublin, National College of Ireland, Sept. 2022. URL: https://norma.ncirl.ie/6106/.

[7] Divyansh. *Dummy Astronomy Data*. https://www.kaggle.com/datasets/divyansh22/dummy-astronomy-data. Accessed: 02/09/2024.

[8] Australia Telescope National Facility. *Main sequence stars*. https://www.atnf.csiro.au/outreach/education/senior/astrophysics/stellarevolutionmainsequence.html. Accessed: 21/09/2024.

[9] GeeksforGeeks. *Neural Networks: A Beginner's Guide*. https://www.geeksforgeeks.org/neural-networks-a-beginners-guide. Accessed: 23/09/2024.

[10] C. Gohd et al. *Stars*. https://universe.nasa.gov/stars/basics/. Accessed: 09/09/2024.

[11] W. Herschel. "Catalogue of a second thousand of new nebula and clusters of stars, with a few introductory remarks on the construction of the heavens (xx)." In: *Philosophical Transactions of the Royal Society of London* 79 (1789), pp. 212–255.

[12] ESA Hubble. *Active galactic nucleus*. https://esahubble.org/wordbank/active-galactic-nucleus/. Accessed: 2024.

[13] HubbleSite. *Galaxies*. https://hubblesite.org/science/galaxies. Accessed: 23/09/2024.

[14] IBM. *Logistic Regression*. https://www.ibm.com/topics/logistic-regression. Accessed: 23/09/2024.

[15] Peng Jia, Qiang Liu, and Yongyang Sun. "Detection and Classification of Astronomical Targets with Deep Neural Networks in Wide-field Small Aperture Telescopes." In: *The Astronomical Journal* 159.5 (Apr. 2020),

p. 212. DOI: 10.3847/1538-3881/ab800a. URL: https://dx.doi.org/10.3847/1538-3881/ab800a.

[16] Dustin Lang and David W. Hogg. *Principled point-source detection in collections of astronomical images*. 2020. arXiv: 2012.15836 [astro-ph.IM]. URL: https://arxiv.org/abs/2012.15836.

[17] Emma Leifer. *Image Classification of Stars and Galaxies Using Different Machine Learning Models*. Oct. 2023. DOI: 10.47611/harp.300.

[18] Yosry Negm. "Astronomical Images Classification Using Deep Learning CNNs." In: *Yosry Negm, Faculty of Electronic Engineering, Menoufia University, Egyp* (February 2021).

[19] Maahi Patel. *The Complete Guide to Image Preprocessing Techniques in Python*. https://medium.com/@maahip1304/the-complete-guide-to-image-preprocessing-techniques-in-python-dca30804550c. Accessed: 06/09/2024.

[20] Towards Data Science. *Why Deep Learning is Needed Over Traditional Machine Learning*. https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063. Accessed: 22/09/2024.

[21] Webb Space Telescope. *What are active galactic nuclei?* https://webbtelescope.org/contents/articles/what-are-active-galactic-nuclei. Mar. 2021.

[22] A Vafaei Sadr et al. "DeepSource: point source detection using deep learning." In: *Monthly Notices of the Royal Astronomical Society* 484.2 (Feb. 2019), pp. 2793–2806. ISSN: 0035-8711. DOI: 10.1093/mnras/stz131. eprint: https://academic.oup.com/mnras/article-pdf/484/2/2793/27689282/stz131.pdf. URL: https://doi.org/10.1093/mnras/stz131.

[23] American Association of Variable Star Observers. *Bl lacertae*. https://www.aavso.org/vsotsbllac. Accessed: 12/09/2024.

[24] Wikipedia. *Quasar*. https://en.wikipedia.org/wiki/Quasar. Accessed: 24/09/2024.