Deep Learning Techniques for Astronomical Object Classification

MSc Research Project
MSc in Data Analytics

Yogiraj Subhash Dalvi
Student ID: 20205741

School of Computing

National College of Ireland

Supervisor: Abubakr Siddig

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Yogiraj Subhash Dalvi		
Student ID:	20205741		
Programme:	MSc in Data Analytics	**Year:**	2021-22
Module:	Research Project		
Supervisor:	Abubakr Siddig		
Submission Due Date:	19th September 2022		
Project Title:	Deep Learning Techniques for Astronomical Objects Classification		
Word Count:	8203	**Page Count:**	20

I hereby certify that the information contained in this (my submission) is information pertaining to the research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other authors' written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Yogiraj Subhash Dalvi

Date: 19/09/2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

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

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

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

Deep Learning Techniques for Astronomical Objects Classification

Yogiraj Subhash Dalvi

20205741

Abstract

As there are an infinite number of deep space objects, it is crucial to differentiate those space objects such as stars, galaxies, or quasars in recent or upcoming deep astronomical surveys. This task becomes very difficult because of the tedious procedure for separation between edgy and expanded sources, which makes this classification problem a difficult task. After Machine Learning approaches, there has been a rise in the use of Deep Learning methodology for deep space object classification challenges because of its improved calibration. Deep learning models such as VGG16, ResNet50, and InceptionV3 were trained using images obtained from the Slone Deep Space Survey. In addition to these pre-built architectural models, we have also implemented a CNN classifier equipped with an adam optimization parameter to explore the behavior of CNN layers. Although this CNN classifier along with other pre-trained models employed in this study was able to extract new features and classify astronomical objects with an accuracy of 79% plus, our VGG16 model achieved a significantly good accuracy of 86.04%.

Keywords*—*Deep Learning, VGG16, Pre-trained models, CNN, astronomical objects

1 Introduction

Image processing in deep learning takes place in many network layers hence, identifying input features and analyzing human-visible images is made possible by hidden layers of artificial neural networks. Many features of data can be extracted using convolutional neural networks. In recent years, many surveys have been carried out and the resulting data have been saved but, understanding all astronomical data is difficult and leads to knowledge loss. As astronomical imagery is captured during the survey, it takes a lot of effort to correctly identify an image. In the future, processing large amounts of data will be resilient and automated as certain types of machines can learn to solve different problems, including classification. The latest developments in machine learning are being used for a variety of tasks such as morphological classification, point source identification, QSO detection, star classification, novelty, and anomaly detection. Convolutional neural networks, known as CNNs, are becoming unwieldy in low-power environments as computing power continues to increase. So usually, people don't know how many layers or filters they need for each layer. Currently, CNN parameters are searched using either heuristics or grid search. Transfer learning is often used by CNN designers when developing new CNNs for new data. This involves adapting the network to specific types of data as one has the option to do a grid search or start with 8 to 64 filters at each level but, this might result in a huge redundant network. Many studies have shown that networks can be reduced in size and can remain reliable but, this task redundancy requires more time, money, and special training. (Garg et al. 2020).

1.1 Motivation

In 2007, astronomer Kevin Schawinski got approximately one million photos of galaxies to examine thanks to the Sloan Digital Sky Survey (SDSS). He believed the experiment might be completed in a more effective manner. To facilitate their work, they conceived of the notion of using armatures, which was made possible thanks to the assistance of Chris Lintott, a fellow at Oxford College as this resulted in the creation of the project known as Galaxy Zoo. People had the misconception that the work would take years to complete, but it was finished in only six months. The SDSS telescope began its operational life in the year 2000 but, since that time, it has been conducting surveys of the night sky. 2019 saw the release of DR16 findings from the Slone survey (Ahumada et al. 2020). They added additional data to previously available data. Up to 2018, the SDSS has images of 1/3rd of the dark sky recorded in 5 different wide bands.

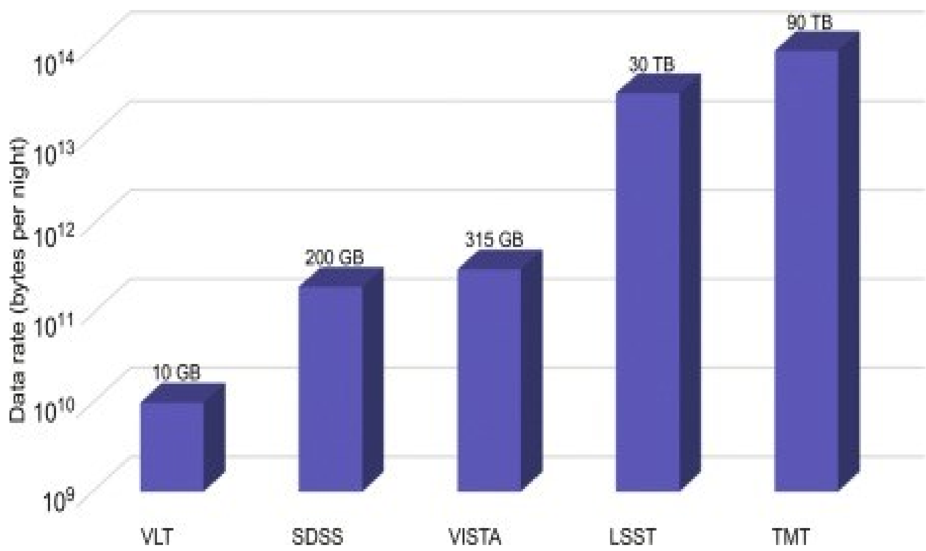

Fig 1. Telescope data rate in Astronomy Imagery (Ahumada et al. 2020)

Astronomy and cosmology are data-rich. As a result, more ambitious sky surveys have been made possible by large aperture digital cameras mounted on telescopes. Now, whole surveys may be gathered in only a single night if necessary. (Lai and Kong, 2020) Figure 1 depicts the amount of data collected by telescopes that are currently operational as well as those that will be constructed soon. These telescopes include the Very Large Telescope (VLT), the

Sloan Digital Sky Survey (SDSS), the Visible and Infrared Telescope for Astronomy (VISTA), the Large Synoptic Survey Telescope (LSST), and the Thirty Meter Telescope (TMT). Digital telescopes of today can collect data at the terabyte level in only one night but, Deep Space image models were a free resource for scientists. Real-world demands need new methods to interpret data as Data scientists assist astronomers to observe astronomical objects with limited resources. To help them in Classifying space objects, I learned about this classification problem out of technical interest.

1.2 Research Question and Objectives

The amount of time you can spend using a highly-priced telescope is limited. In addition to this, it is constrained by the conditions that exist in astronomy and the weather. Because scientists spend most of their time snapping images in deep space and picking out their subjects with great consideration, these experts have a thorough grasp of all that exists in the universe. This study discusses the challenge that astronomers face in the modern world, which is figuring out how to deal with the increasingly large quantity of data that is being obtained, as well as finding ways to enable future telescopes, as well as advancements in astronomical images, and the processing of increasing amounts of data.

Question: *"How can the finest techniques in deep learning for image classification (VGG16, InceptionV3, ResNet50, and CNN based on Adam Optimizer) be employed with the data augmentation technique for classifying images of astronomical objects to aid scientists in astronomical imagery?"*

The following is a list of the primary objectives of this study:

1. Recognition of recent advancements in astronomical image classification through applications of Computer Vision.
2. The use of image classification methods in deep learning for astronomical imaging of space.
3. Assessment of applied models using evaluation metrics.
4. Comparison between model performance and state-of-the-art models.

2 Related Work

Due to a lack of suitable equipment in the late 1990s, the task of seeing a space object in the sky was difficult but, as technology advanced, people were able to study both nearby and distant galaxies. Recently, computerized computer technologies allowed for the morphological analysis and classification of deep space objects. This study describes various studies that used machine learning and deep learning to attempt to solve problems by classifying astronomical images, identifying gaps, and making suggestions for further research.

2.1 Astronomical Imaging

Astronomical image surveys collect spectroscopic or photometric data. Spectroscopic monitors photon wavelengths over hundreds of wavelengths to identify chemical substances like water. Photometry measures just a few broad-band filters using a CCD, resulting in less comprehensive data. There is always a trade-off between the spectroscopic's inability to detect faint or distant things and the expense of measuring fewer items in a single image. Faint objects are farther distant, allowing astronomers to view further back in time. This helps scientists to grasp the origins of life and our existence. Due to artifacts and other distortions, many of the produced images still need human modification and classification. The long exposures may cause digital noise, merging galaxies, or space dust which could cause difficulties for machine learning techniques. Astronomical data classification fell into two categories: conventional approaches, utilizing experts for needing many positive findings in astronomical imagery, or ML techniques, which seem to be a recent subject in this field.

2.2 Machine Learning Techniques

17 galaxies from astronomical photos were classified using Support Vector Machines (SVM). This work examined an automatic galaxy classification method that utilized sparse representation. Feature extraction employed 6 morphological elements: Elongation, Form Factor, Con-vestry, Bounding-rectangle-to-fill-factor, and Bounding-rectangle-to-rectangle-perimeter. (Jenkinson et al. 2014) Bounding-rectangle-to-fill-factor, Elongation, Con-vestry Due to the early nature of the experiment, only the Elliptical, Simple Spiral, and Barred Spiral were studied. 14 galaxies were used to train and evaluate the SVM classifier as no mistakes were made in classifying galaxies. With more galaxies, this value may fall, resulting in more feature variety. (Toschi & Bodenschatz 2019) used a noise-removed image, efficient features, and a support vector machine to accurately classify galaxy images. The testing used 57 Sloan Digital Sky Survey images as a Support Vector Machine was used to classify galaxy photographs based on 827 criteria. In two spiral and elliptic classes and three spiral, elliptic, and zinc-edged classes, accuracy was 96 and 94, respectively. The accuracy for classifying three classes of galaxy images was 97.1%, 97.8%, and 98.2%, and for three classes, it was 94.71%, 95.2%, and 96.33%.

(Xiao-Qing & Jin-Meng 2021) employed machine learning to classify stars, galaxies, and radio waves from LAMOST DR5 (QSOs). The magnitudes u, g, r, I z, J, and H were utilized to classify stars, galaxies, and QSOs. Using any of the four classifiers, it was observed that accuracy exceeded 95% in all situations. The Random Forest achieved the highest accuracy, while KNN placed third. The random forest model correctly predicted 98.9% of the star class data whereas 97.8% of predicted galaxy classifications were accurate. 88% of QSOs were properly predicted as well. (Martinazzo, Espadoto & Hirata 2021) recommended self-supervised learning for astronomical image data, in which a large neural network is pre-trained with unlabelled data and astronomical properties as output, which may be derived from the input photographs alone. This research shows that, even with little quantities of labeled training data, this pre-training boosts accuracy for downstream classification tasks, and that, in most circumstances, the results produced by using Transfer Learning based on ImageNet pre-training are better.

(Machado et al. 2016) compared machine learning techniques for star/galaxy separation using photometric data from astronomy catalogs. In addition to the investigated classifiers, the Xero classifier was introduced, which classifies the dataset based on most classes while ignoring the rest. Trials showed that most studied and explored solutions surpassed Zero in accuracy. It is assumed that a good classification algorithm should beat Zero's accuracy. In this study, both NN and RF did well. Using Hadoop, the exploratory method was a data-centric real-time procedure as it included data selection through classifier performance analysis in one pipeline. The method and parallel implementation are crucial for large-scale data processing and demanding to compute for machine learning technique research. M81 globular cluster candidates were analyzed using methods to pick the most promising astronomical objects. (Chuntama et al., 2021). Using a clustering approach, this study eliminated human classifications and improved labeled data quality. This study shows that data may be sorted into 12 clusters, each of which can be broken into six groups of similar astronomical objects. Using these six data sets, the classification models enhanced their prediction accuracy wherein Multilayer Perceptron enhanced accuracy from 67.1% to 91.59. WIZARD and Random Forest had 90.96% and 90.57% accuracy, respectively. The research shows that the model can classify objects with astronomical properties. This model can't discriminate globular contenders from background galaxies since they're so similar.

(Makhija et al., 2019) used photometric data from the GALEX and SDSS observatories to compare quasars and stars. Despite having similar shapes, the two objects are completely distinct and far apart. These objects with spectroscopic data served as our training set, and photometric sources may benefit from a database of samples classified using our classifiers. This work proposes using a GAN-based classifier to handle classification. Researchers found that the classifiers' accuracy varied from 91% to 100%, which is satisfactory as classifiers may classify samples without spectroscopic labels. (Wang et al., 2019) developed a machine learning model using NIST's Galaxy and Mass Assembly (GAMA) datasets. In addition to the morphological data in SersicCatVIKING and SersicCatUKIDSS, users have access to GaussFitSimple's spectroscopic characteristics, and MagPhys' physical properties, and Lambdar's photometric observations. Five galaxy catalogs were analyzed using Random Forests and GMLVQ, a prototype-based classifier. After reviewing a huge number of additional galaxy features, this research concludes that the visual-based classification technique used to name the galaxy sample isn't supported by the data. However, previous investigations using a lower-dimensional dataset achieved the same result.

RR Lyrae, a notable Milky Way star, was classified by (Singh et al., 2018) using machine learning. Lyrae stars have 0.1-hour daily pulse durations. In this study, variable and non-variable stars were separated using five different supervised learning methodologies which include classification techniques such as the decision tree, support vector machine, logistic regression, Naive Bayes, and AINN. Logistic Regression has the highest accuracy of all techniques, 99.49%. As more space junk is colliding, endangering satellites and space missions in astronomy, it is very important to identify and classify space debris objects to protect space assets. (Khalil et al. 2019) used eight machine learning models to recognize real-world light curves of space objects as they identified space object light curves. FATS, feet, and UPSILON exhibit light curves as feature collections (Automated Classification for Periodic Variable Stars Using Machine Learning). 3.98:96.02 for debris classes and other RSOs affects classification performance. As a result, ADASYN, SMOTE, Borderline-SMOTE, and SVM-SMOTE were rigorously tested. The accuracy, precision, and recall of six distinct classifiers (decision trees, linear discriminant functions, Naive Bayes, SVM, k-NN, and three Ensemble classifiers) were compared. SVM on the FATS feature set produces the most accurate results using Borderline-SMOTE. These findings are 99.1% accurate, 100% precise, and 97.2% recall. SVM-SMOTE and/or Borderline-SMOTE oversampling may provide high classifier results. In this study, all feature sets and oversampling methods were affected.

(Chuntama et al., 2020) classified the astronomical objects seen in M81 into the following five categories: stars, spherical galaxies, elongated galaxies, globular clusters, and fuzzy objects (fuzzy objects). Images obtained from CFTs were processed by the author. Following the visualization of a portion of the data, seven supervised learning techniques were applied to the remaining information. These algorithms included Logistic Regression, SVM, Multiclass Classifier, WiSARD, and Random Forest. For the purposes of building and evaluating classification models, Weka was utilized using the M81 datasets. To evaluate the accuracy of multiclass classification models, 10-fold cross-validation was used. On a total of eight confusion matrices, accuracy, precision, recall, and F-measures were measured and computed. Random Forest outperformed the other six multiclass classification models with an accuracy score of 81.2%, a precision score of 81.0%, a recall score of 81.2%, and an F-measure of 80.6%.

(Guzman et al., 2018) published a 15-experiment automatic stellar star classification approach in Astronomy Astrophysics. This research showed that with proper setup, Classifier Systems may achieve excellent classification accuracy. To increase classification, several factors were evaluated. Calculating the suggested parameters uses Chebyshev Coefficients, Fourier, Wavelet, and Comb Moments. This work includes the design of Stellar Spectral Classification utilizing the Harvard System, using five and seven classes, and the capture of the spectral image. This article describes a strategy for classifying stars using astronomy images. The SVM classifier classified 5 of 7 classes with 88.1% accuracy. K-NN had the highest accuracy, 90.32 percent when considering the seven spectral classes. (Du Buisson, et al., 2015) used SDSS photos to study supernovas and image artifacts. Humans still remove noise and artifacts from images. Using 8 Eigen image features (PCA of single-epoch g, r, and I difference images), 96% recall was obtained. Random forests, k-nearest neighbor, and Skynet (Graff et al. 2014) ANN algorithms outperformed naive Bayes and kernel SVMs.

2.3 Image Classification and Deep Learning

Computer vision and signal processing leverage deep machine learning. This research classified images using deep learning classifiers like Convolutional Neural Networks (CNN). CNN allows a machine to learn complicated visual features from its depiction, minimizing human knowledge. (Khalifa et al. 2018) utilized three categories to classify galaxies which are irregular, spiral, and elliptical. While classifying these galaxies, their properties were considered. An eight-layer convolutional neural network architecture was used in this study. This approach includes a feature extraction layer with 96 convolutional filters as well as two classification layers that are entirely interconnected. The architecture was subsequently improved with the aid of 4238 additional images used as training data. Image enhancement techniques such as rotation, reflection, cropping, and Gaussian noise, to name a few, were used in the training data to increase image quality. Deep Galaxy V2 is an effort to overcome the overfitting problem by introducing an augmentation process to the training data. The purpose of this method is to decrease the probability that the model will overfit.

(Ren et al. 2016) present a model of convolutional neural network initialization for image classification based on principle component analysis. This paper advises employing principal component analysis to get eigenvectors and initialize convolutional kernels without supervision. Gradient diffusion is less of an issue with adequate starting parameters since they include image information. All this research suggests it's possible to enhance classification accuracy while limiting iterations, optimizations, and training time. (Patil et al. 2021) studied how residual neural networks (ResNet-18, ResNet34, ResNet50, and ResNet152) recognize galaxies. This research shows that ResNet-152 outperforms other residual networks and helps determine galaxy classification trustworthiness. Adding residual networks improved the model's accuracy. ResNet152 outperformed the other two models. Reset models outperformed other models on the dataset.

Convolutional neural networks have won several competitions in recent years as it works great on Image recognition. Edward used ConvNets to classify stars and galaxies based on reduced, calibrated pixel values. This research utilized data from the Sloan Digital Sky Survey and the Canada-France-Hawaii Telescope Lensing Survey to illustrate ConvNets' efficacy. (Huertas-Company et al., 2021) provided a summary of the historic and current impacts of using machine learning to predict galaxies' sizes and forms. Supervised CNN-based classifications were accurate for identifying huge samples of galaxies, according to this study. Transfer learning methodologies or even generated datasets may be a beneficial alternative to extensively labeled datasets in many instances, such as image recognition. Numerical simulations and deep neural networks may be utilized to examine physically driven classifications. Despite being successful, supervised learning can't discover new things. This research examined unattended and supervised medical techniques. Large imaging surveys may be used to identify and examine outliers and compare observed and simulated galaxy populations.

(Mohamed Selim et al., 2022) presented a modular strategy for automatically detecting a galaxy's optical canter, area, and classification. A test on 1000 galaxies from EFIGI proved the research's validity. Sharpness is adjusted to rectify faint galaxies, then the noise is removed as the approach proceeds by examining galaxies' visual centers. The visual canter is used to find sections of galaxies that provide classification information wherein galaxy brightness fluctuations are used to classify galaxies. The classification of galaxies is 97.2% accurate and takes 0.37 seconds per galaxy on average. Galaxies may be hard to discern because of their darkness, vivid background stars, and image noise. An innovative, modular approach suggested by (Essa et al., 2022) uses raw brightness data to estimate a galaxy's visual canter, region, and classification. First, a new selective brightness threshold is employed to make galaxy visual centers easier to detect with brilliant background stars. This is the second way to identify galactic regions. A new approach to identifying galaxies considers how their brightness changes over time. The researcher tried this approach on 1000 EFIGI galaxies which yielded good results. In 0.37 seconds, 97.2 percent of galaxies were analyzed. High success rates and fast processing times showed that the work was executed efficiently.

2.4 Identified Gaps

Telescopes produce more data than people can manually analyze. Traditional techniques of classifying objects, such as Machine Learning techniques, have a purpose, but they can't keep up with the volumes of data or added complexity as deep space survey technology evolves. While neural networks aren't new, they are a relatively recent addition to astronomers' toolkits. However, the intricacy of developing models and classifying necessary to train a model has only lately been publicly accessible. The following section shows that Deep Learning methods like ConvNets can be used to identify astronomical objects like

Stars, Galaxies, Quasars, etc. TensorFlow, an open-source machine learning platform for computer vision recognition applications, which now supports Keras, a python-based open-source neural network toolbox for deploying deep learning models. Upon its latest upgrades, it is clearly providing an opportunity to our use for using it in astronomical imagery to classify objects.

3 Research Methodology

While working on this project, several different approaches to data analysis were investigated, such as KDD, CRISP-DM, and SEMMA. After giving a thought to it, it was concluded that the best course of action would be to stick with the KDD approach since the individual was already knowledgeable about it. In addition, KDD is centered on the process of deriving knowledge from data within the setting of huge databases, which has been one of the key base points for this study. Hence, KDD was adapted for use in astronomical imagery.

3.1 KDD Methodology for Astronomy

The project's research was carried out in a methodical and professional way by adhering to the KDD Methodology. Each of the following subsections examines how the project performed each of the phases within this technique, as specified in mentioned, as well as how each step was customized to meet the requirements of the research being conducted for this project. The first thing that needed to be done was to acquire knowledge of the topic at hand, assess the level of previous information, and determine the objectives for this study. This topic is discussed in the next sections, where a critical analysis of this research is presented.

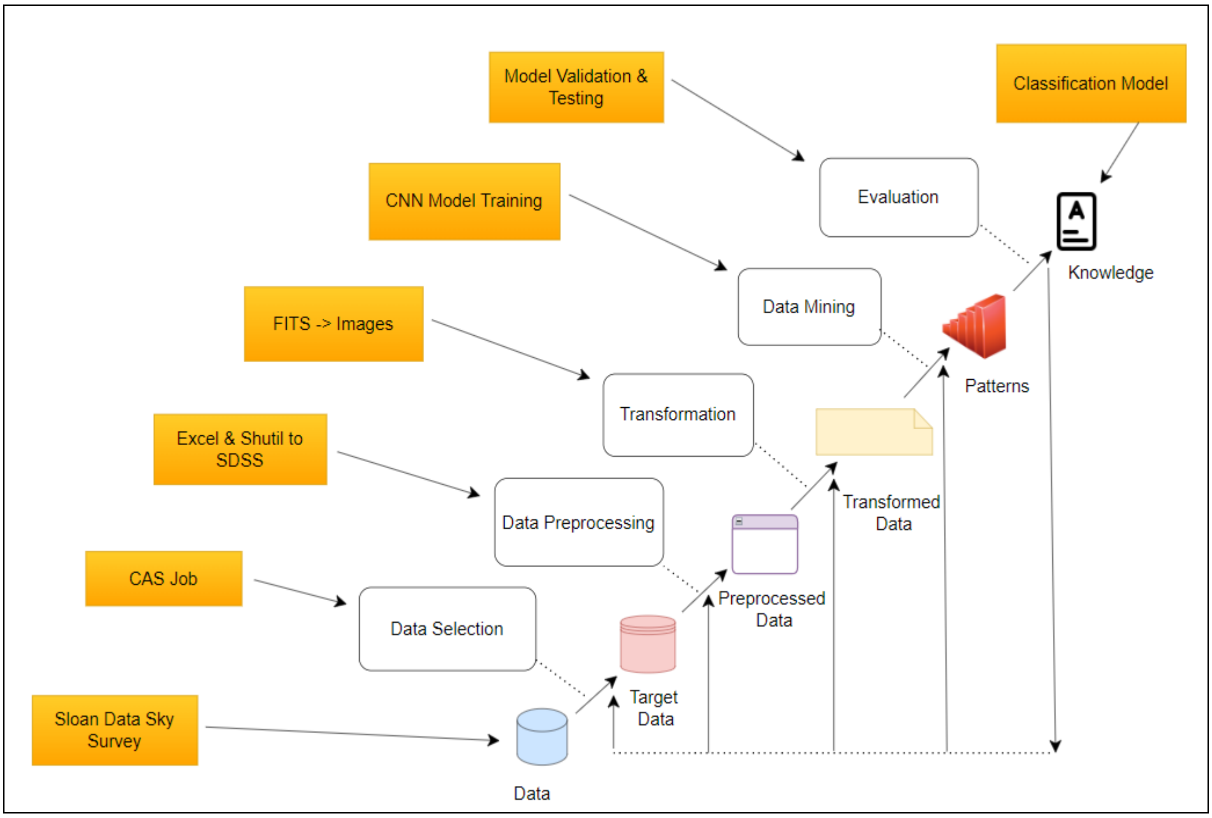

Fig 2. KDD process for Astronomical Object Classification

3.1.1 Data Selection

FITS is the name of the format that is most often used for the gathering of astronomical data. With the help of the Flexible Image Convey System, one can convey not only data but also image data (FITS). On the other hand, photographs used in astronomy are often captured in grayscale with several filters in place to pick up various wavelengths. After then, these images may be put together to provide a full spectrum image of the thing under question.

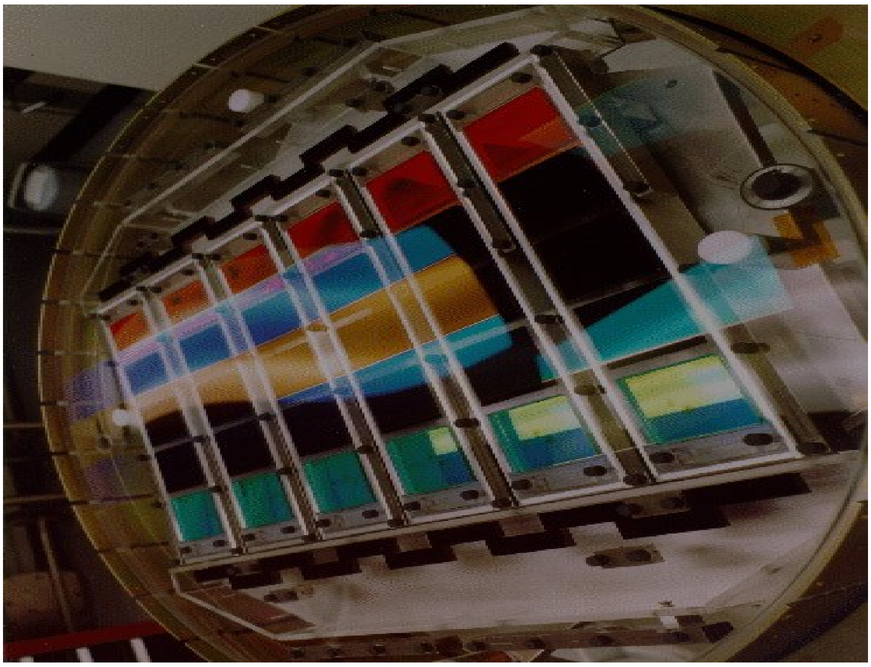

Fig 3. SDSS Imaging Camera (Sdss.org., 2022)

The above figure shows an imaging camera that may follow five filters: ultraviolet (UV), green (g), red (both visible and infrared), and infrared. When carrying out a CAS job on the SDSS Catalogue Archive Server, an online SQL query was executed to fetch data. To get started, we utilized the data from the archive to locate all the objects and then, classify them according to their Right Ascension and Declination (RA and DEC) which includes classifying them as either galaxy (like our own Milky Way) or stars or quasars (like Sun). The dataset for this study includes images from Data Release 17 (Sdss.org, 2022.).

3.1.2 Data Pre-Processing and Transformation

Following the division of the objects into 3 separate classes, a query was executed on SDSS to determine the number of instances of each object class. The 3000 objects from the CAS job were selected to represent each SDSS category. This equated to five filters being applied to each FITS image for over 9000 images. The images were then randomly divided into training datasets and testing datasets. This was to ensure that each FITS image could utilize all available filters. The "cut-out" jpegs from the FITS files were then retrieved locally via a web API request that was made to the SD17 JPEG web service. All the images were resized to 224 by 224 pixels, and after that, the training images were augmented by rotating, shifting, shearing, and flipping the images to create some more data for training. The imaging data has been rescaled from [0,255] to [0,1] to normalize the input for processing.

3.1.3 Data Mining

Data mining extracts valuable data from unstructured data. The objects were identified by classification from catalog files and extracted into acceptable forms for modeling. It includes data analysis to find trend lines, data collecting, and occasionally data processing into data warehousing systems. To mine the data and extract properties for the purpose of classifying astronomical objects, convolutional neural networks were used. In this study, the following neural network architectures were utilized: VGG16, InceptionV3, ResNet50, and CNN based on Adam Optimizer. The method of learning consisted of loading the training weights for the models that had been trained using ImageNet, removing the topmost layer, and adding new layers to adapt the models to new image classes.

3.1.4 Data Evaluation

For the evaluation of this project, the four important metrics have been considered which are: Accuracy, Precision, Recall, and Confusion Matrix. Accuracy is how many right predictions a model has produced. The confusion matrix y model measures the number of correctly and incorrectly predicted test records. The confusion matrix shows which classes are correctly anticipated and which are incorrectly forecasted, as well as the types of errors made. Precision displays how many predictions were accurate. Recall shows how many positive instances our model properly predicted. This project evaluated whether computer vision and classification methods might help identify and classify astronomical imagery. The recorded results are presented in the next evaluation section.

4 Design Specification

Computer vision has advanced significantly over the last few years, outside of cloud and mainframe computing environments, however, its ability to scale has been challenging to achieve because of its complex design and many connected weights. However, Deep Learning techniques and convolutional neural networks have helped to change this scenario. Now, we will try to understand the process flow for this study.

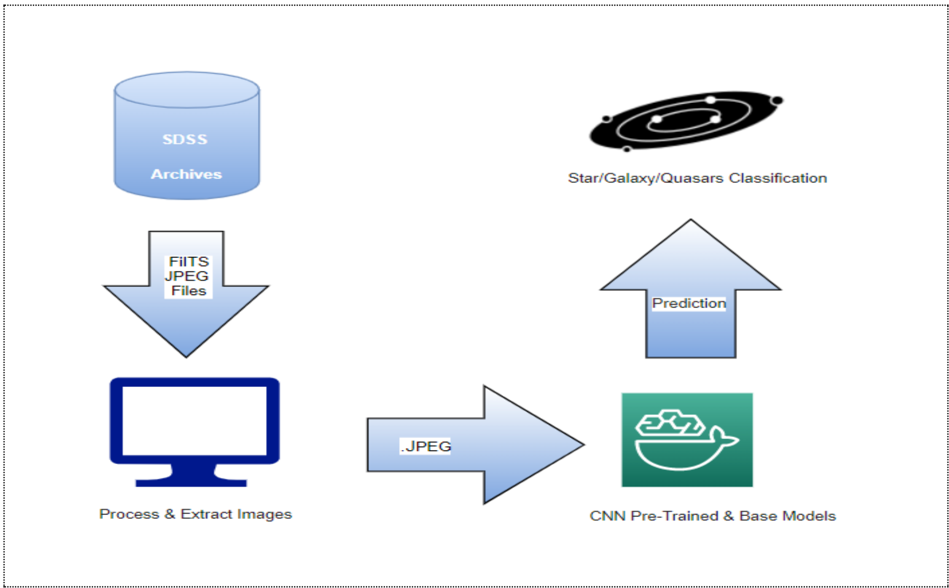

Fig 4. Process Flow for Astronomical object classification

4.1 Base Architecture for Convolutional Neural Networks

Convolutional Neural Network is one of the types of neural network which is used in Deep Learning. Image recognition has taken a significant leap forward thanks to CNN's work. CNNs include many layers, including an input layer, an output layer, and hidden layers. These layers all contribute to the processing and classification of images in some way. All layers are included in the hidden layers. Each of these layers plays an important part in the overall network.

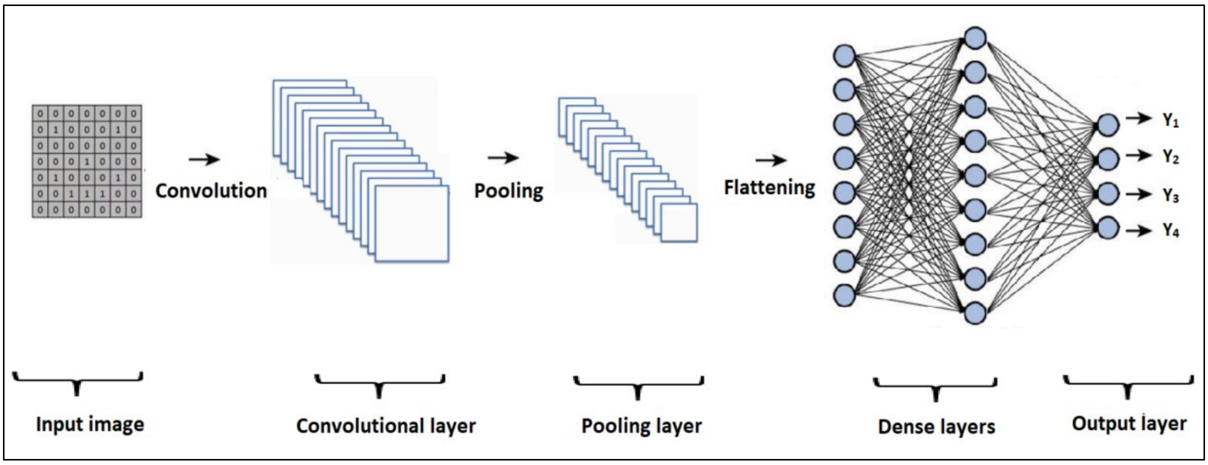

Fig 5. An architecture of Convolutional Neural Network (Mdpi 2022)

4.2 TensorFlow

An open-source framework for machine learning was developed by the Google Brain Team and given the name TensorFlow. Because of a team of engineers and resources provided by both the TensorFlow group and Google, using TensorFlow is now within the reach of most individuals who do not have a solid understanding of the mathematics that underlies the operation of ANNs. (TensorFlow., 2022) As a result of the various degrees of abstraction provided by the framework, users are free to focus on the implementation rather than the code that underpins the functions. The support was provided for a significant number of computer languages, the most notable of which being Python.

4.3 Data Acquisition

The Sloan Digital Sky Survey (Sdss.org., 2022.) is an imaging space research project that collects data on the night sky by using a 2.5-meter wide-angle telescope at the Apache Point Observatory in New Mexico. This observatory is in the United States. On a regular basis, the dataset is made accessible to the public. The dataset for this study contains information from DR17. (Sdss.org. Data Release 17., 2022) The information was collected with the help of SDSS catalog servers by means of a python script that retrieved image data through an online SQL interface. When retrieving the data using a SQL query, the galaxy, star, and Quasar classes were taken into consideration. To download the image files, another Python script was executed, and this time it used the URLs of the FITS files as its input. The files that were downloaded were sorted according to the class to which each image belonged.

4.4 Data Processing and Modelling

A Python script was used to process the FITS files to extract the images and save them as individualc.jpg files, one for each filter. All JPEG images were retrieved with the help of a customized script that was made available as an API to the data servers of that website. After some time, they were saved locally from the personal space that SDSS computing allots to each of us after creating an account in it.

4.5 Presentation of Data

The results of the models have been shown in the form of five different evaluation metrics: Accuracy, Precision, F1 Score, and Classification Report. The training accuracy, validation accuracy, Training loss function, and Validation loss function have been considered while considering the overall model's performance. To measure how well or poorly the model is working, the loss function is used.

5 Implementation of CNN Classification Models

The following section shows how the KDD process has been implemented in this study.

5.1 Data Selection

On the SDSS Catalogue Archive Server, also known as CAS, a SQL query was executed, which resulted in an output that was in CSV format. This provided a listing of all attributes that were afterward applied to the process of identifying the necessary FITS files and the JPEG file that was downloaded from the SDSS server. An example of the output is included as well as the techniques that were used to compile the results of the query into URLs that permitted image selection and download. Another Python script was run, and this time it took as its input the URLs of the FITS files. This allowed the image files to be downloaded successfully. The downloaded files were organized into classes that corresponded to the kind of images that were being retrieved.

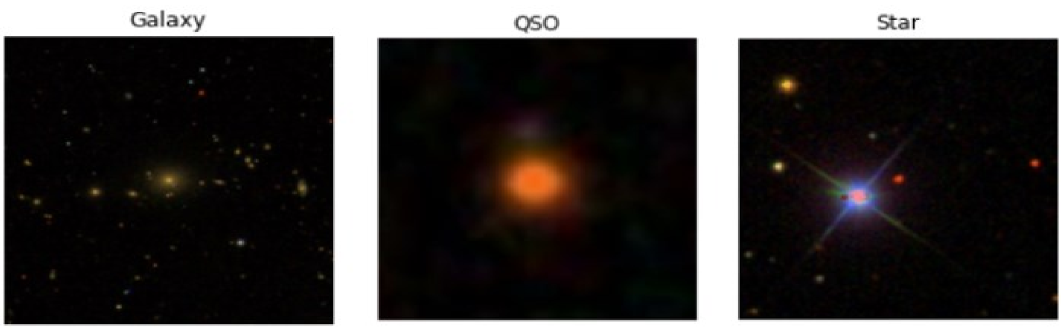

Fig 5. Sample Images from SDSS Data Release 17

5.2 Data Pre-Processing

After the images were separated into three distinct classes, a query was run on SDSS to fetch the images for each object class. The 3000 images that were part of the CAS job were chosen to be representative of each SDSS object class. This meant that each FITS image had five filters applied to it. After that, the images were sorted into training datasets and testing datasets. This was done to ensure that each FITS image could make use of all the filters that were accessible. Following that, the cut-out jpegs from the FITS files were downloaded locally by making a web API request to the SD17 JPEG web service. Below table 1 shows the total files downloaded for each class. It includes file type and Image source information as well.

Table 1: File Count by each class

Source	File Format	Galaxy	Quasar	Star	Total
Sloan Data Sky Survey (SDSS)	JPEG	3090	3042	3060	9192

Each image file was given a name that corresponded to the class that it was placed in so that it could be identified more easily in the future. After the FITS files had been downloaded, the URLs were removed from the pathway, and the file itself was the only thing that remained. This was done so that the files could be managed more easily in the future.

5.2.1 Image Augmentation

Image data augmentation is a technique that modifies the original images that are included in a training dataset to produce new versions of those images. This makes it possible to artificially expand the size of the dataset without the need to acquire any more data in the process. The capacity of fit models to generalize what they have learnt to apply to new images may be improved through augmentation approaches, which allow for the creation of different versions of the images. Training deep learning neural network models on more data may result in more proficient models and training these models can increase their capacity to generalize what they have learnt to apply to new images. There is a vast range of image processing available, including the following:

Image Flipping: This will result in the image being flipped either horizontally or vertically.

Image Rotating: This will allow the image to rotate by a certain angle, either clockwise or counterclockwise, depending on the direction you choose.

Rescaling: The rescale value 1./255 will convert the pixels that are in the range [0,255] will be converted to the range [0,1] for processing. This can be also considered Input Normalization.

Image Cropping: In this process, a part of the image is selected at the random for cropping.

5.3 Data Mining

The CNN data mining process starts with the extraction of features. Following that is a flattening layer, followed by an adjustable number of ReLU layers, and finally, a SoftMax layer. Finding the convolutional and ReLu layers requires more imagination than hard scientific work According to the findings of (Ma, Dang, and Li 2014) study, the number of hidden layers used for image identification is determined not by science but rather by a process of trial and error. ImageNet (ImageNet., 2022) is an image database structured by the

WordNet hierarchy, with thousands of images for each node. The Keras application has several deep learning models embedded in it. They all produced model performance results in comparison to the ImageNet dataset (Table 2), in which accuracies correspond to the model's results in comparison to that dataset.

Table 2: Keras Application Models Accuracy on ImageNet dataset (Team, K., 2022)

Model	Size (MB)	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth	Time (ms) per inference step (CPU)	Time (ms) per inference step (GPU)
Xception	88	79.0%	94.5%	22.9M	81	109.4	8.1
VGG16	528	71.3%	90.1%	138.4M	16	69.5	4.2
VGG19	549	71.3%	90.0%	143.7M	19	84.8	4.4
ResNet50	98	74.9%	92.1%	25.6M	107	58.2	4.6
ResNet50V2	98	76.0%	93.0%	25.6M	103	45.6	4.4
ResNet101	171	76.4%	92.8%	44.7M	209	89.6	5.2
ResNet101V2	171	77.2%	93.8%	44.7M	205	72.7	5.4
ResNet152	232	76.6%	93.1%	60.4M	311	127.4	6.5
ResNet152V2	232	78.0%	94.2%	60.4M	307	107.5	6.6
InceptionV3	92	77.9%	93.7%	23.9M	189	42.2	6.9
InceptionResNetV2	215	80.3%	95.3%	55.9M	449	130.2	10.0
MobileNet	16	70.4%	89.5%	4.3M	55	22.6	3.4

In our methodology, we have used some of the models from the above table as a top layer of our CNN models, which classified images when trained against the ImageNet dataset, were removed in our models, instead, the weights of ImageNet were implied and all the layers of the model were frozen, meaning initial patrameters were untrainable, then the new layers were added to classify our new astronomical images.

5.3.1 Fine-Tuning of CNN Models:

Transfer learning includes aspects such as fine-tuning as one of its important components. This technique involves unfreezing parts of the top layers of a model that had been frozen in the past and then jointly training the new top layer, which is used to classify our astronomical image datasets, together with the remaining layers of the frozen model. We are doing this because the top-level feature maps can only be extracted by the very final layers of the base model, whereas the initial convolution layers can only extract the most fundamental features such as edges, vertical lines, horizontal lines, etc. This fine-tuning of the top feature representation in the model enables us to make them more specific for the classification problem that we are attempting to solve. Even when using identical settings for all the hyperparameters such as learning rate, batch size, etc., using various random seeds might provide very different outcomes. The problem is considerably more obvious, particularly when using the huge variations of Transfer learning on little datasets, which is where it really stands out.

6 Evaluation and Discussion

This section examines the results of all the models that were examined, compares their performance with the images from SDSS and then evaluates their results to the performance of models that are state of the art.

6.1 VGG16 Model

This model was constructed by (Simonyan and Zisserman, 2014) which aimed to study the accuracy of CNN’s on massive image datasets. Their study looked at the potential that the accuracy of the CNN may be improved by using convolutional filters that were extremely tiny (3x3) and by increasing the number of weighted layers that were included inside the CNN itself. This model overall performed well on astronomical image classification.

	precision	recall	f1-score	support
Galaxy	0.88	0.98	0.93	307
QSO	0.78	0.88	0.83	302
Star	0.96	0.71	0.82	304
accuracy			0.86	913
macro avg	0.87	0.86	0.86	913
weighted avg	0.87	0.86	0.86	913

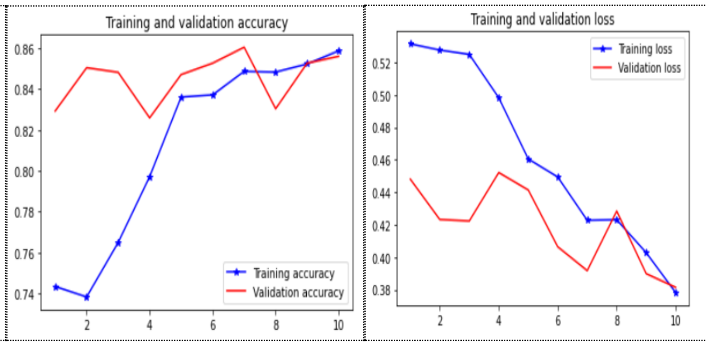

Fig 6. Classification Report and Plots for VGG16 Model (After Fine-Tuning)

Figure 6 shows the classification report and Training-Validation plots for the VGG16 model. It achieved 86.04% accuracy in the classification of SDSS imagery. The F1 scores in the classification report show that the galaxy images shared more success in classification than that Stars and Quasars. Towards the conclusion of the 10 epochs, there was a drop in validation accuracy, which may be something that needs to be examined in further study. The loss function for both training and validation followed a downward trend as it is reaching towards 10 epochs. The training and validation loss function almost reached zero at the end of the 10 epochs.

6.2 InceptionV3 Model

On the ImageNet dataset, the image recognition model known as Inception v3 achieved an accuracy of 77.09 percent. The model is a synthesis of the findings of many researchers over an extended period. It is based on the original work that was written by (Szegedy et al. 2016) titled "*Rethinking the Inception Architecture for Computer Vision.*" Figure 7 shows the classification report and Training-Validation plots of the InceptionV3 model on our SDSS imagery. It achieved 83.92% accuracy in the classification of SDSS imagery. The galaxy image classification was more successful than the image classification of Stars and Quasars. Yet, the processing done on the SDSS images showed a solid increasing trend in the accuracy, and the model has not attained a 0 loss.

Classification Report

	precision	recall	f1-score	support
Galaxy	0.87	0.97	0.91	307
QSO	0.79	0.78	0.79	302
Star	0.87	0.77	0.82	304
accuracy			0.84	913
macro avg	0.84	0.84	0.84	913
weighted avg	0.84	0.84	0.84	913

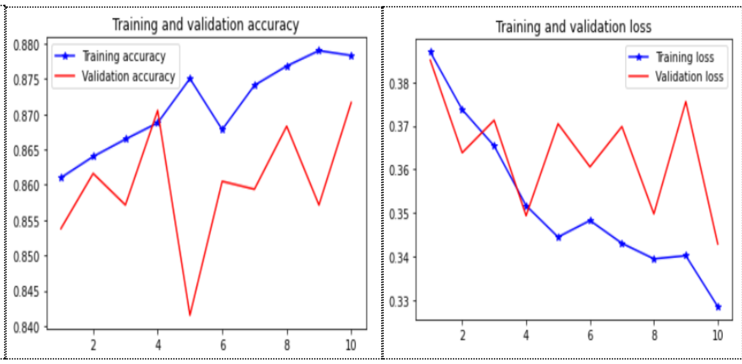

Fig 7. Classification Report and Training-Validation plots for InceptionV3 Model (After Fine-Tuning)

As a finding, there was the further possibility to enhance this result with more training that was observed. The model demonstrates the ability to continue learning with more epochs, and it is a candidate for future study and development to obtain a higher level of precision.

6.3 ResNet50 Model

The ResNet506 model, which was developed by (He et al. 2016), consists of layers that are broken up into blocks and includes over 23,000,000 trainable parameters. The deterioration and the disappearing gradient problem were two of the challenges that were associated with extremely deep learning, and it is credited that ResNet50 was able to overcome these challenges. When it came to processing SDSS files, the model also performed at a level comparable to that of other models. The fact that it included an extra one million parameters that could be tuned as part of the training did not result in a substantial increase in the outcomes.

Classification Report

	precision	recall	f1-score	support
Galaxy	0.83	0.97	0.89	307
QSO	0.69	0.87	0.77	302
Star	0.98	0.55	0.70	304
accuracy			0.80	913
macro avg	0.83	0.80	0.79	913
weighted avg	0.83	0.80	0.79	913

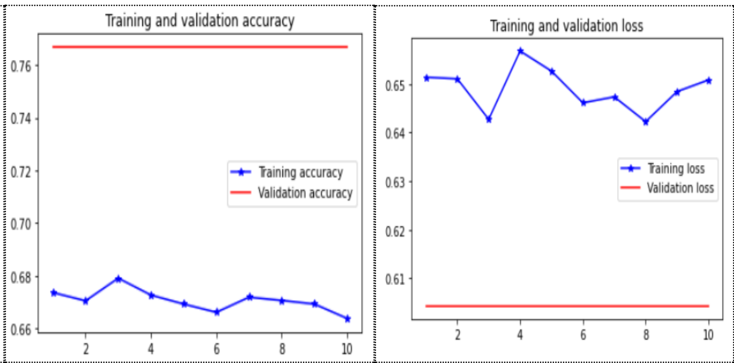

Fig8. Classification Report and Training-Validation plots for ResNet50 Model (After Fine-Tuning)

Figure 8 shows the classification report and Training-Validation plots of the ResNet50 model on our SDSS imagery. It achieved 79.79% accuracy in the classification of SDSS imagery. The model demonstrates the capability of continuing to improve with more epochs, and it is potential for future work and for attaining a higher level of precision. The F1 score valued for Quasars came out low compared to other classes. The validation accuracy and loss function part of the ResNet50 model did not improve at all between 10 epochs, could be a part where input data might need to be increased hence, needs thorough investigation for in future work. It did not demonstrate the same capability as InceptionV3 to keep honing its learning and become better at its task.

6.4 Base Convolutional Neural Network with adam optimizer

This model uses identity mapping and an Adam optimizer to learn interrelations and extract image features. (Wang et al., 2022) Four identity mappings enable deep layers to directly learn data from subsurface layers, reducing gradient vanishing caused by network depth. A novel Adam optimizer with a power-exponential learning rate is presented to regulate CNN iteration direction and step size. CNN model is trained using an exponential Adam optimizer to speed up its optimization. The power-exponential learning rate controls iteration direction and step size to fast achieve the ideal solution.

Classification Report

	precision	recall	f1-score	support
Galaxy	0.83	0.97	0.89	307
QSO	0.69	0.87	0.77	302
Star	0.98	0.55	0.70	304
accuracy			0.80	913
macro avg	0.83	0.80	0.79	913
weighted avg	0.83	0.80	0.79	913

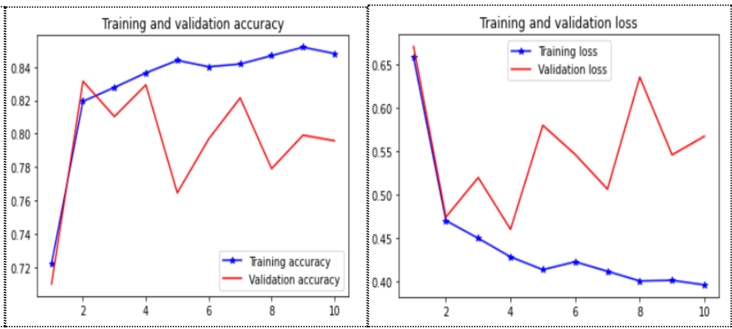

Fig 9. Classification Report and Training-Validation plots for CNN with adam optimizer (After FineTuning)

Figure 9 shows the classification report and Training-Validation plots of the base CNN model with adam optimizer. It achieved 79.57% accuracy on SDSS space object classification. The training accuracy of this model followed an upward trend however, validation accuracy kept fluctuating by the end of epochs. Like other pre-trained models, galaxy-class classification for this model achieved a better F1 score than Quasar and Star classes.

6.5 Review of Implemented Models

All the models had the exact same modifications carried out to them, such as the removal of the final output layer that was going to be used to display the outcomes of the classification performed using the ImageNet images and the insertion of the same number of extra layers in each model. This enables a direct comparison of apples to apples in terms of the performance of the various models. When it comes to the amount of time it took to process, size did matter with this training.

The datasets that were used in this study are separated into three different classes, and training and validation images in the count from 3,000 to 5,000 were applied to the models. However, the ImageNet dataset was used in the past to train the models that were implemented here. This is a dataset consisting of 14.5 million photos that have been classified into little over 21,500 categories. We are considering these model accuracies against ImageNet dataset as our baseline accuracy for comparison. The below table discusses the results of all implemented models with respect to training, validation accuracies and other factors. All the models were fed with the pre-trained parameters that were obtained from the training on the ImageNet dataset. This allowed models to have information about extracting image features. These findings are equivalent to those that were obtained by ImageNet's training, which can be seen in Table 3.

Table 3. Performance Review of Implemented Models

Implemented Model	VGG16	InceptionV3	ResNet50	CNN with adam optimizer
Performance Against SDSS Imagery				
Total Layers	19+4	313+4	50+4	12+4
Total Trainable Params	1,28,47,107	21,05,347	63,07,843	2,21,76,803
Total Non-Trainable Params	1,47,14,688	2,18,06,880	2,35,97,952	0
Epochs	10			
Training Accuracy	**0.8675**	0.8465	0.8172	0.8480
Validation Accuracy	**0.8604**	0.8392	0.7979	0.7957
Performance Against State-of-the-Art Models (ImageNet Dataset)				
Accuracy	0.7130	0.7790	0.7490	x

In comparison with the other four pre-trained CNN architectures that have been built, it is obvious that VGG16 performed exceptionally well.

7 Conclusion and Future Work

The objective of this project was to study the performances of different Pre-trained CNN architectures for astronomical object classification and to determine if they could be implemented in this field to help scientists classify objects in astronomical imagery.

This study has implemented four different architectures of CNN models however, it has been observed that VGG16 achieved higher training and validation accuracies than other pre-trained CNN models. Another important thing to observe in this study is that the Stars and Quasars classes achieved fewer F1 scores compared to the Galaxy class. This could be an interesting task in future work. However, this could be because Stars and Quasars have some similar features in nature.

The GPU-based TensorFlow processing helped at some points to lower training times. Pre-trained layers of CNN architectures allowed google colab to process data quickly. 10 epochs were implemented on each model's performance to create a baseline for comparison.

This study addressed the classification of three classes only but, as an improvement, it should include more than three classes as there could be thousands of different astronomical objects as well, such as supernovas. It could be also interesting to see if we can introduce Image Segmentation techniques in this classification task as SDSS FITS files covers too much area and have too many objects for classification in a single FITS file. Also, TensorFlow's high processing GPU power might help astronomers to detect and classify objects through live video streaming of astronomical objects.

8 Acknowledgement

I would like to express my gratitude to my supervisor, Prof. Abubakr Siddig, and his assistance throughout the research.
I would also like to thank my parents and my friends and classmates for their support and encouragement while doing research and writing my thesis. Without them, this achievement would not have been possible.
Lastly, I would like to express my gratitude to the experts and support team who were involved in the Sloan Data Sky Survey. This research would not have been able to be carried out effectively in the absence of their ardent involvement and contribution. Also, to the team at TensorFlow for providing a licensed open-source platform.

References

Ahumada, R. and Anders, F., 2020. The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. *The Astrophysical Journal Supplement Series*, 249(1).

Bastanfard, A. and Amirkhani, D., 2019, October. Automatic Classification of Galaxies Based on SVM. In 2019 9th International Conference on Computer and Knowledge Engineering (ICKE) (pp. 32-39). IEEE.

Chuntama, T., Suwannajak, C., Techa-Angkoon, P., Panyangam, B. and Tanakul, N., 2021, June. Classification of Astronomical Objects in the Galaxy M81 using Machine Learning Techniques II. An Application of Clustering in Data Pre-processing. In 2021 18th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 1-6). IEEE.

Chuntama, T., Techa-Angkoon, P., Suwannajak, C., Panyangam, B. and Tanakul, N., 2020, December. Multiclass Classification of Astronomical Objects in the Galaxy M81 using Machine Learning Techniques. In 2020 24th International Computer Science and Engineering Conference (ICSEC) (pp. 1-6). IEEE.

Du Buisson, L., Sivanandam, N., Bassett, B.A. and Smith, M., 2015. Machine learning classification of SDSS transient survey images. Monthly Notices of the Royal Astronomical Society, 454(2), pp.2026-2038.

Eassa, M., Selim, I.M., Dabour, W. and Elkafrawy, P., 2022. Automated detection and classification of galaxies based on their brightness patterns. Alexandria Engineering Journal, 61(2), pp.1145-1158.

Garg, I., Panda, P. and Roy, K., 2019. A low-effort approach to structured CNN design using PCA. IEEE Access, 8, pp.1347-1360.

Guzmán, A.P.A., Esquivel, A.E.O., Hernández, R.D. and Robles, L.A., 2018, October. Stellar Spectral Classification with Machine Learning Classifier Systems and Chebyshev Coefficients using the Tonantzintla Schmidt Camera Plate Collection. (pp. 81-87). IEEE.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Jenkinson, J., Grigoryan, A., Hajinoroozi, M., Hernández, R.D., Barreto, H.P., Esquivel, A.O., Altamirano, L. and Chavushyan, V., 2014, October. Machine learning and image processing in astronomy with sparse data sets. In 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 200-203). IEEE.

Khalifa, N.E., Taha, M.H., Hassanien, A.E. and Selim, I., 2018, March. Deep galaxy V2: Robust deep convolutional neural networks for galaxy morphology classifications. In 2018 International Conference on Computing Sciences and Engineering (ICCSE) (pp. 1-6). IEEE.

Khalil, M., Fantino, E. and Liatsis, P., 2019, December. Evaluation of Oversampling Strategies in Machine Learning for Space Debris Detection. In 2019 IEEE International Conference on Imaging Systems and Techniques (IST) (pp. 1-6). IEEE.

Lai, J. and Kong, L., 2020. Big Data in Astronomy. ScienceDirect, pp.29-58.

Machado, E., Serqueira, M., Ogasawara, E., Ogando, R., Maia, M.A., da Costa, L.N., Campisano, R., Guedes, G.P. and Bezerra, E., 2016, July. Exploring machine learning methods for the star/galaxy separation problem. (IJCNN) (pp. 123-130). IEEE.

Makhija, S., Saha, S., Basak, S. and Das, M., 2019. Separating stars from quasars: Machine learning investigation using photometric data. Astronomy and Computing, 29, p.100313.

Martinazzo, A., Espadoto, M. and Hirata, N.S., 2021, January. Self-supervised Learning for Astronomical Image Classification. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 4176-4182). IEEE.

Patil, J.S., Pawase, R.S. and Dandawate, Y.H., 2017, May. Classification of low-resolution astronomical images using convolutional neural networks. (pp. 1168-1172). IEEE.

Ren, X.D., Guo, H.N., He, G.C., Xu, X., Di, C. and Li, S.H., 2016, June. Convolutional neural network based on principal component analysis initialization for image classification. (pp. 329-334). IEEE.

Sdss.org. 2022 | SDSS Imaging Camera | SDSS. Available at:<https://www.sdss.org/faceplat/>

Sdss.org. 2022. Data Release 17 | SDSS. Available at: <https://www.sdss.org/dr17/>

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Singh, J., Joshi, G., Tiwari, H., and Tiwari, I., 2018, August. Applied Machine Learning for and Analysis of RR Lyrae Variable Stars. 2018 2nd International Conference on (pp. 247-252). IEEE.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

Team, K., 2022. Keras documentation: Keras Applications. [online] Keras.io. Available at: <https://keras.io/api/applications/>.

TensorFlow. 2022. TensorFlow. [online] Available at: <https://www.tensorflow.org/>

Xiao-Qing, W. and Jin-Meng, Y., 2021. Classification of star/galaxy/QSO and star spectral types from LAMOST data release 5 with machine learning approaches. Chinese Journal of Physics, 69, pp.303-311.

Y. Wang, Z. Xiao, and G. Cao, “A convolutional neural network method based on Adam optimizer with power-exponential learning rate for bearing fault diagnosis,” Journal of Vibroengineering, Vol. 24, No. 4, pp. 666–678, Jun. 2022