Combining Clustering and Classification methods for Galaxy Morphology Identification

MSc Research Project
Data Analytics

Ian Dias
Student ID: 22205748

School of Computing
National College of Ireland

Supervisor: Dr. Paul Stynes, Prof. Musfira Jilani

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ian Dias
Student ID:	22205748
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Dr. Paul Stynes, Prof. Musfira Jilani
Submission Due Date:	12/08/2024
Project Title:	Combining Clustering and Classification methods for Galaxy Morphology Identification
Word Count:	5285
Page Count:	21

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

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

Signature:	Ian Dias
Date:	14th September 2024

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 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):	

Combining Clustering and Classification methods for Galaxy Morphology Identification

Ian Dias
22205748

Abstract

Galaxy Morphology is the study of galaxies based on their shapes and structures. Traditional research primarily uses classification or clustering techniques to categorize galaxies into distinct groups such as Elliptical and Spiral Galaxies. Currently as more telescopic surveys are planned to be launched, the challenge that is faced by astronomical scientists, is to classify these huge amounts of data for further research, although supervised techniques do work well, scientists have shown concern when working with human labelled data due to potential biases. Thus, this research proposes a new machine learning framework to potentially reduce human annotations in data by combining Image Classification and Clustering techniques. The framework combines a Hierarchical Clustering technique using the Entropy, Gini and Gradient Moment (EGG) coefficients, with Neural Networks. The research will conduct two tests, by implementing the Hierarchical Clustering using HDBSCAN, paired with the classification model EfficientNetB0, and for the second test combining Self Organising Maps along with a CNN architecture. The SDSS catalog containing approximately 670,000 galaxy jpeg images and FITS data, out of which approximately 13,452 of both, comprising nearly 50% of each class, will be used to conduct this research, pertaining to the limited availability of resources and time constraints. The results show the hdbscan+efficientNetb0 and Som+CNN frameworks giving an accuracy of 90% and 93% respectively.

Keywords: Galaxy Morphology, Classification, Clustering, HDBSCAN, EfficientNetB0, SOM, CNN

1 Introduction

With an introduction of new and larger Telescopic Surveys coming up, such as the Sloan Digital Sky Survey (SDSS), there is a need for astronomical researchers and data scientists to turn to machine learning models to classify galaxies and identify new ones and group them into categories based on their shape and sizes such as Elliptical and Smooth galaxies Fraix-Burnet (2023); Lahav et al. (1995); Zhu et al. (2019). Handling SDSS data is challenging due to its vast size and complexity, hence there is ongoing research to develop robust models that helps automate classification. Savić et al. (2023) Current advancements in this field include using supervised classifications in the form of neural network architectures such as CNN, VGGNet, VGG16, ResNet, EfficientNetB0, etc.Baumstark and Vinci (2024); Becker et al. (2021); Chen (2023); Kalvankar et al. (2021); Patel (2023) and supervised classification using traditional machine learning methods such as SVM's, Random Forest etc. Baumstark and Vinci (2024); Savić et al. (2023) and unsupervised methods like hierarchical clustering models Cheng et al. (2021); Rosito et al. (2023); Yu and Hou (2022) to classify galaxies based on their morphological characteristics. There is also ongoing research on using a combination of supervised and unsupervised approaches, as supervised approaches require labeled data which could be time-consuming and prone to human biases Kolesnikov et al. (2024); Dai et al. (2023); Ma et al. (2023). Due to an increase in the amount of data generated due to telescopic surveys and with the possibility of discovering new galaxy structures, there is a need to develop more scalable, efficient, and robust models, which can contribute further to our understanding of the universe.

The purpose of the research is to build an efficient and scalable framework that can be used on the data collected by the SDSS and answer the question **"How well can the combination of clustering models and classification models reduce need for human annotations in the morphological classification of galaxies?"**. This research proposes a new framework that combines the highly optimized and efficient CNN architecture in the form of EfficientNetB0 with the robustness and scalability of the HDBScan architecture, as mentioned in the papers Kalvankar et al. (2021); Nazeri (2023); Tan and Le (2020). The research will use the sdss data in the form of fits files and jpg images requested from Igor Kolesnikov, which can also be obtained from the sdss website, used in papers by Kolesnikov and Dominguez Sanchez Domínguez Sánchez et al. (2018); Kolesnikov et al. (2024). The formed clusters will also be evaluated and checked whether or not they align with already existing classifications and whether there are new or unexpected morphological characteristics.

This paper discusses the use of clustering methods to form clusters from image data and then use these formed clusters to train neural networks to create a framework that can classify galaxy data without need of pre-existing labels. Related work in this field is discussed in Section 2. The research methodology is discussed in section 3 followed by design specification and implementation in sections 4 and 5, the results and evaluations are discussed in section 6. Section 7 concludes the paper and discusses future work.

2 Related Work

Classifying galaxies play an important role in helping astrophysicists and astronomical data scientists understand the universe and its expansion and identifying any peculiarities in galaxy structure. In recent years, as astronomical surveys have become more advanced and vaster in size, it becomes a near impossible task to accurately and reliably classify galaxies with the help of human participants due to there being biases involved Fraix-Burnet (2023), Fraix also points out the sensitivity of the algorithms towards 'features present in images,' which would not necessarily coincide with Hubble classifications [1]. Lahav et al. Lahav et al. (1995) was among the first to integrate machine learning in morphological classifications and have also highlighted the importance of using deep learning for this task in their research.

The work of Domínguez Sánchez et al. (2018) introduces a morphological catalog for a large number of galaxies of the Sloan Digital Sky Survey using CNN. The authors addressed the issue of color bias of the galaxies by removing the colour information

[1] Hubble Classifications: `https://mcdonaldobservatory.org/sites/default/files/pdfs/teachers/HubbleClassificationSheet.pdf`

from the image catalog; this was done so that the classification depends entirely on the morphological characteristics, i.e. the shape and size of the galaxy. This prepared catalogue has been made available by her and have been since used in research Kolesnikov et al. (2024), which is further used by the author in this research.

2.1 Supervised Learning for Galaxy Morphology Identification

Most common deep learning approaches have been the use of deep learning models such as CNN Becker et al. (2021); Domínguez Sánchez et al. (2018); Zhu et al. (2019); Lukic et al. (2018); Shi et al. (2023); Tang et al. (2022), RNN Fielding et al. (2021); Patel (2023); Zhu et al. (2019); Gupta et al. (2022); Wu et al. (2022), ANN(Lahav et al. (1995), VGG architectures Chen (2023); Patel (2023); Zhu et al. (2019), DenseNet Fielding et al. (2021) and EfficientNet architecturesFielding et al. (2021); Kalvankar et al. (2021); Wu et al. (2022), etc. and traditional machine learning methods such as random forest classifier, support vector machines, xtreme gradient boosting and decision trees Baumstark and Vinci (2024); Savić et al. (2023). All of these methods provide good results and are able to capture the hidden features successfully. But as these rely on pre-defined labels, scientists wanted to look for a way to segregate galaxies without human biases.

2.2 Unsupervised Learning for Galaxy Morphology Identification

Thus in order to try and eradicate human biases, clustering techniques for galaxy classification are being used as they do not need labelled data to work with. This approach relies on the calculated metrics obtained from FITS file data generated by telescopic surveys[2]. Commonly used metrics are the Concentration (C), Asymmetry (A), Clumpiness (S), Entropy (E), Gini (G), Gradient Moment (G2) Conselice (2003); Kolesnikov et al. (2024); Cheng et al. (2021); Baumstark and Vinci (2024). By using clustering techniques, the model becomes less dependent on labels and is based solely on the calculation of algorithms to distinguish groups. Most effective clustering methods with respect to galaxy classification are found to be the Hierarchical clustering methods Cheng et al. (2021); Ma et al. (2023); Rosito et al. (2023); Yu and Hou (2022). These methodologies are able to remove human biases and segregate galaxies based on metrics. Now the other challenge was that clustering algorithms could not be run every time new data came in, thus a combination of unsupervised and supervised learning came into picture, where the basic ideology is to use the cluster labels and train the supervised models with them.

2.3 Hybrid Unsupervised-Supervised Learning for Galaxy Morphology Identification

As mentioned above, a combination of unsupervised and supervised approaches are being researched on Dai et al. (2023); Kolesnikov et al. (2024); Wu et al. (2022). These methods generally include using the clustering methods to form groups and then train the classification model using those groups to form a framework. This addresses both of the previously mentioned challenges and is capable of creating a flow to train models using clustering.

[2] SDSS: `https://www.sdss.org/`

Research Gap: In the studies carried out by Domínguez Sánchez et al. (2018); Kalvankar et al. (2021); Wu et al. (2022), classification does provide good results, but all these studies point out to the problem of there being human biases in the classification of galaxies.

Therefore, a solution researchers came up with was to use clustering algorithms in an attempt to group galaxies based on their morphological characteristics and remove human biases Rosito et al. (2023); Yu and Hou (2022). Although these algorithms work fine, depending solely on clustering models was not feasible.

So, further a combination of supervised and unsupervised approaches have been come into consideration Dai et al. (2023); Kolesnikov et al. (2024). The aim of these research's is to create a model able to group and train classifiers based on the identified groups.

The current state of the art is the usage of combination of SOMBrero and CNN proposed by Kolesnikov et al. (2024) which is run on the R language and currently does not have a python implementation. From the above discussed topics there is still need for research on more robust and scalable models and creating a framework to automate classifications. This research proposes a new framework which uses HDBSCAN techniques to cluster galaxy data and then train EfficientNetB0 on these formed clusters.

3 Methodology

The research methodology consists of two different methodology structures namely the Supervised methodology and the Combined Methodology which consists of a framework combining the unsupervised and supervised architectures as shown in Fig.1 and Fig.2 respectively.

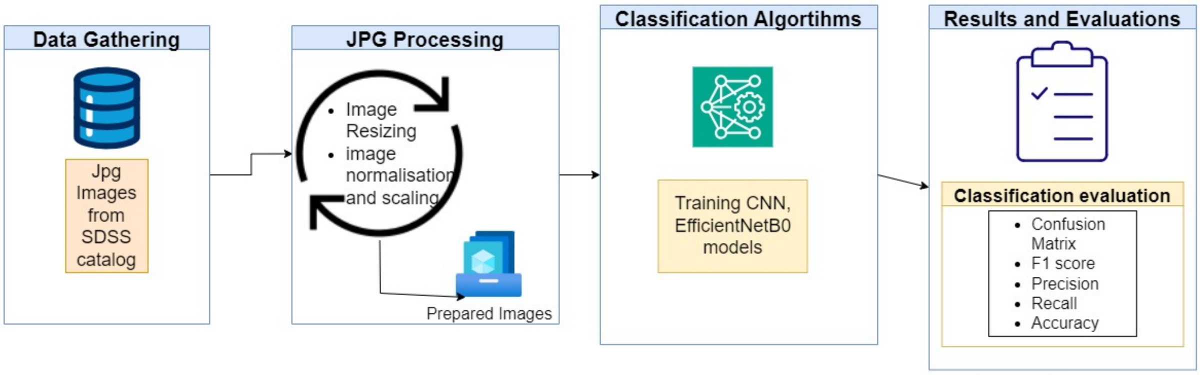
Data Gathering
Jpg Images from SDSS catalog
JPG Processing
Image Resizing
image normalisation and scaling
Prepared Images
Classification Algortihms
Training CNN, EfficientNetB0 models
Results and Evaluations
Classification evaluation
Confusion Matrix
F1 score
Precision
Recall
Accuracy

Figure 1: A diagram giving an overview of the **Supervised learning methodology** followed in this paper

The **data gathering** step remains common for both methodologies, as both data formats can be acquired from the same website. The Combined Methodology has an extra pathway of the clustering segment to generate labels which involves using FITS data for calculating metrics and further next generating labels.

The chosen methodologies and machine learning models are inspired by the works of Domínguez Sánchez et al. (2018); Becker et al. (2021); Wu et al. (2022); Kolesnikov et al. (2024); Cheng et al. (2021), with experiment 6.1 being based on the works of Mr. Kolesnikov.

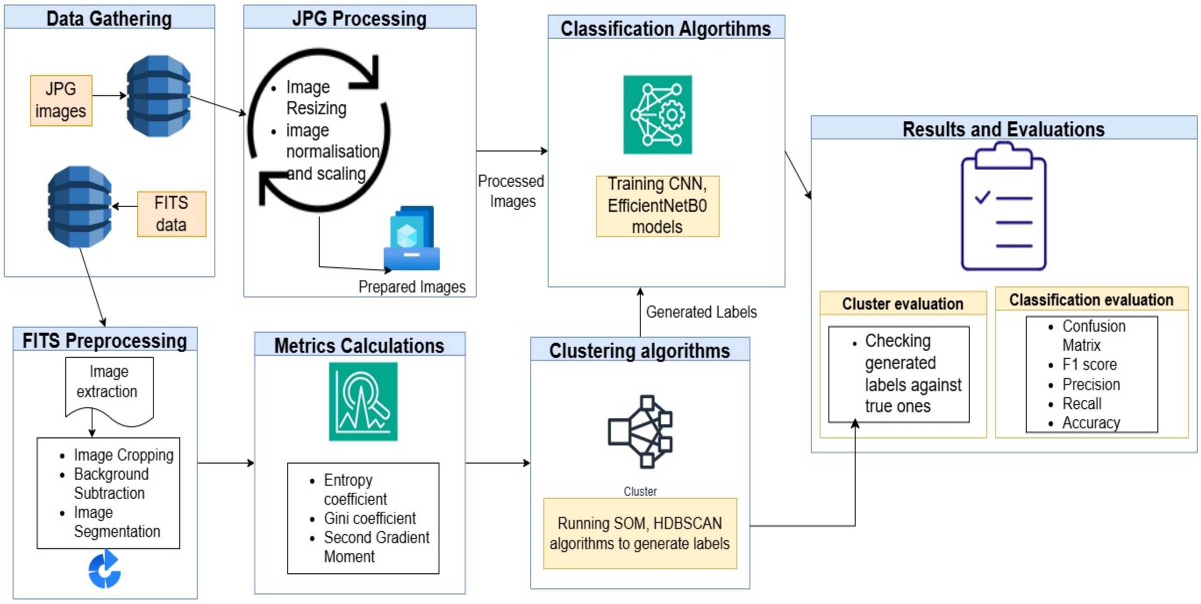

Figure 2: Diagram of the proposed **Combined Methodology** used in this paper, The above branch shows the supervised classification steps and the below branch shows the generation of labels using the clustering algorithms.

3.1 Data Gathering

The first step, Data Gathering involves importing data from the sdss catalogue fits file and corresponding image files of the same galaxies in the jpeg file format.Domínguez Sánchez et al. (2018); Kolesnikov et al. (2024).For this research, the data has been requested by the author from Mr. Igor Kolesnikov, a researcher from the paper Kolesnikov et al. (2024). FITS data is File transfer format commonly used in astronomical research due to its ability to store multiple data formats in the same file and store more bits per pixel in images, further discussion on this and processing is provided in section 3.2 and 3.2.1

3.2 Data Processing

The second step in the flow is processing the respective datasets to feed into the algorithms. The reason for using FITS format file for the initial metric calculations is that these file structures are able to store more bits per pixel for the same galaxy image which helps in getting more accurate measures of luminosity, color, and morphology of galaxies, this helps the unsupervised algorithms define clusters more accurately than having them run only on the jpeg or png format images.Cheng et al. (2021); Kolesnikov et al. (2024) [3]. For the supervised algorithms, using the jpeg format image is sufficient, and doing so will prevent unnecessary processing of the extra bits in the images.

3.2.1 Processing FITS data

For the purpose of this study, only the photometric data has been extracted from the Fits files, further the data is scaled down to cutout of 100 x 100 pixels focusing only on the galaxy in the image file and removing any secondary astral formations, this is done

[3] Astronomical data types: `https://voyages.sdss.org/preflight/capturing-recording-light/types-of-data/`

so that all the further metric calculations remain consistent. Next, the cutout is further cleaned by creating a primary mask and removing that from the original cutout so that the remaining astral formations are removed. The primary mask creation is as below Eq 1.

$$P(i,j) = \begin{cases} \text{True} & \text{if } C(i,j) > \text{median}(C) \\ \text{False} & \text{otherwise} \end{cases} \tag{1}$$

Here P is the Primary mask matrix and C is the cutout matrix, This is used to get a clean cutout, on which the next step is to apply Gaussian Smoothing in order to smoothen the data and remove noise. Kolesnikov et al. (2024); Lotz et al. (2004) [4]. The next and potentially the most important step is to generate the segmentation map of the image data, this is done using the 'SEP' Python library [5]. It involves three steps, detecting background of the image, subtracting it from the smoothed cutout and then creating a segmentation mask and segmented image, which means that all pixels that do not belong to the galaxy are assigned to 0, this can be seen in Fig.3 and Fig.4

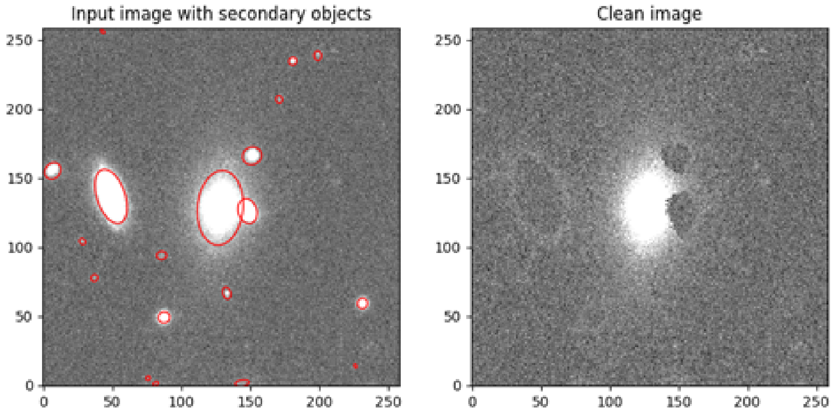

Figure 3: Original vs Cleaned image: shows the capturing of the objects in the image and removing all except the galaxy to segment.

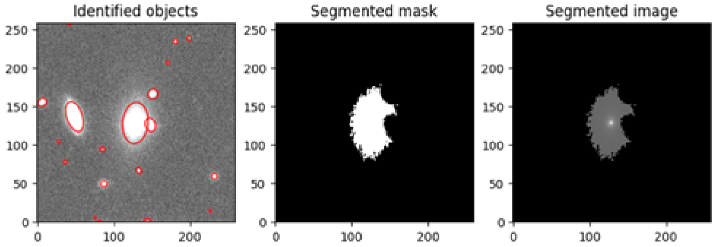

Figure 4: Image, Mask and Segmented Image: Phases of getting an segmented image out of the raw image.

3.2.2 Processing Jpeg Images

Processing the jpg image files involve three simple steps, the first step is to scale the images to the model specific dimensions, the step in which the fits files are scaled to

[4] Gaussian Smoothing: `https://tinyurl.com/45ukcb66`

[5] SEP library: `https://sep.readthedocs.io/en/v1.1.x/tutorial.html`

100x100 pixels has no effect on this, for the CNN architecture the images are scaled to 128x128 pixels and for the EfficientNetB0 architecture the images are scaled to 224x224 pixels respectively. The next step is to convert the images into an array format and normalize the images within the scale 0 and 1. To address the issue of data imbalance, approximately **7000 images of both classes** are taken for training and metrics calculations, by running a Python code to pick a 7000 fits file and the images with the same names of eliptical and spiral classes, respectively.

3.3 Metrics Calculations

This step involves calculating the EGG metrics of the galaxies, namely the Entropy, Gradient pattern analysis (second moment) and Gini coefficients. These metrics are calculated only for the Combined Methodology, to be fed into the clustering algorithm. The three metrics are explained in Sections 3.3.1, 3.3.2, and 3.3.3. CyMorph along with custom functions have been used for calculations [6].

3.3.1 Entropy Coefficient

The Entropy Coefficient, originating from Shannon entropy, measures the distribution of pixel values in an image and captures the randomness in the image's information content. It basically calculates the concentration of pixels by evaluating the pixel density across the specified **number of bins**, which for this research shall be **130**, this helps analyse how the flux distribution is divided. Mathematically, entropy $E(I)$ for variable I is calculated as follows

$$E(I) = -\sum_{k} p(I_k)\log[p(I_k)] \tag{2}$$

Where, $p(I_k)$ as the likelihood of the occurrence of the value I_k, k denotes a specific value and K being the total bin count. Elliptical galaxies generally tend to exhibit lower entropy than spiral galaxies; this can be linked to elliptical galaxies having naturally smoother flux distributions than spiral galaxies. Bishop (2006); Cheng et al. (2021); Ferrari et al. (2015); Kolesnikov et al. (2024).

3.3.2 Gini Coefficient

The Gini coefficient (G) which is originally used in the field of economics to represent wealth distribution has recently started being used for galaxy morphology to measure the relative flux distributions across the pixels, it measures the concentration index, and increases the with an increase in light fraction in the image. Although the Gini coefficient correlates with concentration, it does not presume that the center is the location of the brightest pixel. G is defined mathematically as follows.

$$G = \frac{1}{2\bar{X}n(n-1)} \sum_{i=1}^{n}\sum_{j=1}^{n} |X_i - X_j| \tag{3}$$

where n denotes the galaxy's pixel count and Xi means the flux value of the ith pixel. Elliptical galaxies often display high G values implying that the overall light emission of the galaxy resides in a few pixels, while spiral galaxies tend to have low G

[6] CyMorph: `https://cymorph.readthedocs.io/en/latest/metrics.html`

values suggesting a more even light distribution across the pixels.Abraham et al. (2003); Kolesnikov et al. (2024); Lotz et al. (2004).

3.3.3 Gradient Pattern (Second Moment)

The Gradient Pattern Analysis generally comprises of four moments, out of which only the first two are relevant to galaxy morphology Sautter (2018). For this research, the second moment Gradient pattern (G_2) is used as defined in Rosa et al. (2018). The extraction process of G_2 involves identifying pairs of pixels at equal distances from the center of the image and comparing their modulus (strength) and phase (direction). This step involves calculating two phases, confluence, and then finally the G_2 coefficient. The calculations are shown in Eq. 4 and 5, respectively.

$$\text{confluence} = \left(\frac{\sum_i V_A v_a^i}{\sum_i V_A |v_a^i|} \right) \tag{4}$$

$$G_2 = \frac{V_A}{V - V_c} (2 - \text{confluence}) \tag{5}$$

where, va represents the list of asymmetrical vectors, V_A denotes the count of asymmetric vectors. Spiral galaxies generally tend to have elevated G_2 values due to their disturbed gradient field as opposed to elliptical galaxies. Cheng et al. (2021); Kolesnikov et al. (2024); Rosa et al. (2018); Rosito et al. (2023); Sautter (2018)

3.4 Model Training

This research has two separate methodologies, one comprising of the simple supervised learning methodology, which is done so as to set a baseline for the new framework and see how well it works in comparison. The second methodology is the Combined Methodology which relies on clustering to generate labels for training the classification algorithms. An overview of both methodologies is discussed below in Sections 3.4.1 and 3.4.2

3.4.1 Supervised Methodology

The first set of experiments will be conducted using the CNN and EfficientNetB0 architectures for the classification algorithms following the works of Domínguez Sánchez et al. (2018); Wu et al. (2022). These experiments are carried out to compare the proposed hybrid framework to the plain classification models and see how close can relying solely on metrics for label generation can come to defined labels.

3.4.2 Combined Methodology

The Combined Methodology consists of two experiments involving a combination of Self-Organizing Maps (SOM) and Convolutional Neural Networks and another approach of Hierarchical Density Based Scan (Hdbscan) and EfficientNetB0. More details on these selected methods are mentioned in the following sections. The optimal clustering algorithms which work on astronomical data as such have been proven to be the hierarchical clustering algorithms Kolesnikov et al. (2024); Ma et al. (2023); Rosa et al. (2018); Yu and Hou (2022). Following the clustering steps, based on the generated labels, the neural network models are further trained. In doing so, a framework is developed in such a way that the models are trained without the need of training labels, thus avoiding human biases in the data labelling process. Savić et al. (2023)

3.5 Evaluating Results

The next step is to run evaluations. Two separate sets of evaluations will be run for the separate machine learning algorithms, the clustering and the supervised learning algorithms, and both evaluation steps will be discussed in the following sections.

3.5.1 Clustering Evaluations

In order to evaluate the performance of the clustering algorithms in correctly classifying the galaxies, the labels generated by the clustering algorithms will be compared to the original labels in order to get an idea of how well the clustering algorithms perform in segregating the two classes based on the metrics, and a percentage of accurately formed clusters will be calculated based on how many galaxies are placed into the right cluster.

3.5.2 Classification Evaluations

For evaluating the classification results, the confusion Matrix along with the Precision, Recall and F1 score of the models shall be calculated.

4 Design Specification

This section will discuss the architecture and parameters of the models being used. Sections 4.1 and 4.2 discuss classification and clustering models, respectively.

4.1 Classification Models

The classification models that will be used are the CNN and EfficientNetB0 architectures, as mentioned in section 3.

4.1.1 Convolutional Neural Networks (CNN)

The data processing steps for this architecture is as mentioned in section 3.2.2. The architecture used is shown in fig 5. A dropout layer of 0.5 and a label smoothing of 0.1 have been added to prevent overfitting. The loss function applied is **Categorical Cross-Entropy**, and the optimizer is **Adam**.

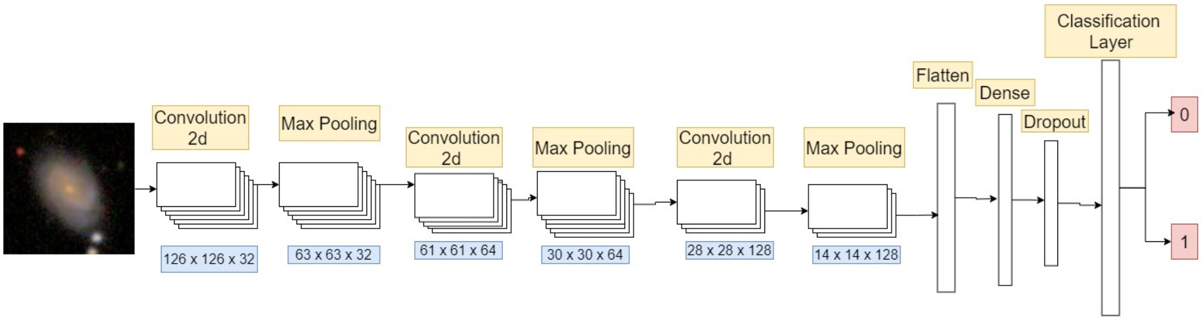

Figure 5: CNN architecture used for this research

4.1.2 EfficientNetB0

Another classification model used for the research is the EfficientNetB0 and its architecture is shown in fig 6. The data processing steps are the same with the only exception being the pixel size as explained in sec 3.2.2. the **Binary Cross-entropy** Loss Function is used with the **Adam optimizer**.

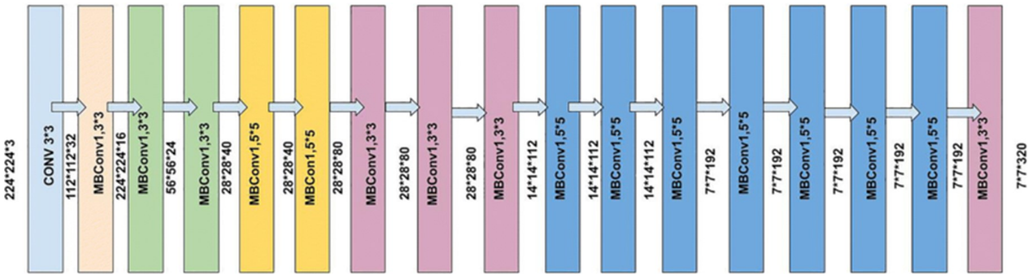

Figure 6: EfficientNetB0 architecture

4.2 Clustering Models

The clustering models that will be used are the Self-Organizing Maps (SOM) and Hierarchy Based Density Scan (HDBScan) architectures.

4.2.1 Self-Organizing Maps (SOM)

The SOM architecture has been used for the clustering aspect of the first combination; the clusters are formed based on the extracted metrics calculated in section 3.3. The bimodality or opposite characteristics of the calculated metrics show a clear difference for the formation of two major clusters, which will be the "elliptical" and "spiral" galaxy classes. The SOM used in the first experiment was chosen to test the effect of centroid based hierarchical clustering and because of the good visualisations of its feature map. The optimal grid size for the algorithm is selected using the formula $grid\_size = \sqrt{(n\_subjects * 0.1)}$ Yaa et al. (2023). The **sigma** parameter and **learning rate** are set to **0.3** and **0.005** respectively. The numeric inputs to the algorithm were scaled and transformed to a range of 0 to 1.

4.2.2 Hierarchical density based Spatial Clustering of Application with Noise (HDBScan)

The second clustering algorithm is HDBSCAN, this is used to test the performance of a density-based hierarchical clustering algorithm as opposed to that of a centroid based clustering algorithm. Hdbscan also has good visualizations in the form of a tree plot that helps to understand the hierarchy of the data points. The inputs for passing into the algorithm are scaled and transformed in the same way as in section 4.2.1 and the **minimum cluster size** is set to **16**, the λ value of **5.0** is taken so as to keep a moderate distance relation between the points in a cluster.

5 Implementation

For the implementation of this project the use of Jupyter Notebooks on **Google Collab Pro** has been done, using the **TPU V2 with high RAM** enabled. This is essential as the data processing may take up high amounts of RAM. The storage of data has been done in **Google Drive**

For the Supervised Methodologies, the images need to be labelled properly by either using a predefined table or labelling the images based on the folders their stored in. This project makes use of **13,452 images** with and **80/20 train test split** for model training.

For the Combined methodologies, it is essential that the same FITS data files are chosen for the same galaxy id images. This can be done on the local system using jupyter notebook, by simply finding files with matching name and storing them for further training. Another important step is to discard any galaxy data that gives a **null or nan metric** value as that may affect the clustering.

The libraries that will need to be installed are, **numpy, pandas, astropy.io, math, scipy.signal, sep, cymorph, astroquery.sdss, scipy.stats, skimage.measure, scipy.ndimage, minisom, hdbscan, sklearn.preprocessing, matplotlib, pylab, sklearn.cluster, scipy.spatial.distance, scipy.spatial.distance, scipy.cluster.hierarchy, tensorflow, sklearn.model_selection, tensorflow.keras.preprocessing.image, pickle, tensorflow.keras.models, tensorflow.keras.losses, tensorflow.keras.layers, sklearn.metrics, tensorflow.keras.applications**.

6 Evaluation

This section shall demonstrate the findings of experiments 1 through 4 in the following sections 6.1, 6.2, 6.3 and 6.4 and then furthermore compare the findings and discuss the results in section 6.5.

6.1 Experiment 1 (CNN)

The first experiment is the application of the Convolutional Neural Network for the classification of galaxies based on their morphological characteristics. The CNN architecture is one of the best working algorithms for this purpose, as stated in the works of Domínguez Sánchez et al. (2018); Becker et al. (2021).

After carrying out the experiment according to the specifications discussed in the 4 section, figure 7 displays the confusion matrix of the results.

The Confusion Matrix displays that the CNN algorithm is able to classify galaxies very well with an accuracy of close to a 100% and the model accuracy and loss charts display the training phase across 30 epochs.

6.2 Experiment 2 (EfficientNetB0)

The second experiment follows the works of Kalvankar et al. (2021); Wu et al. (2022) and another reason to carry out this experiment was the study of Tan and Le (2020).

Fig 9 and 10, show the results of the experiment in the form of the confusion matrix and the model accuracy and loss charts.

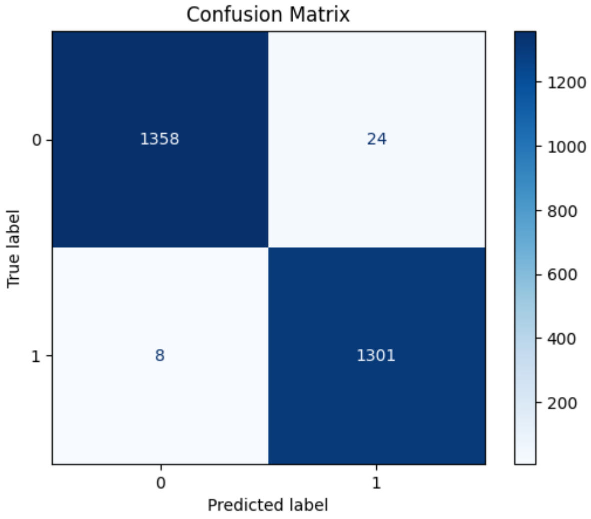

Figure 7: CNN confusion Matrix: O's indicate spiral galaxies and 1's indicate eliptical galaxies

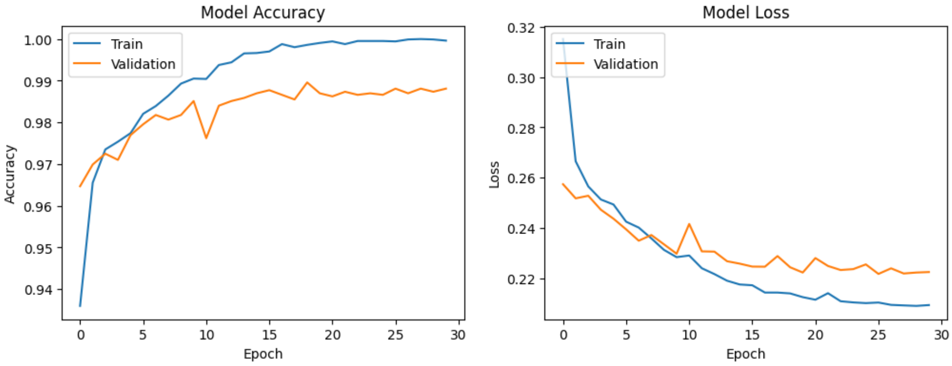

Figure 8: CNN Model Accuracy and loss charts

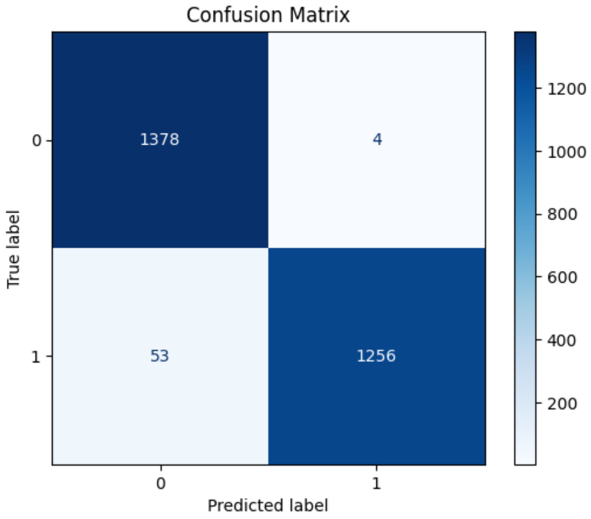

Figure 9: EfficientNet Confusion Matrix: 0's indicating spiral galaxies and 1's indicating eliptical galaxies.

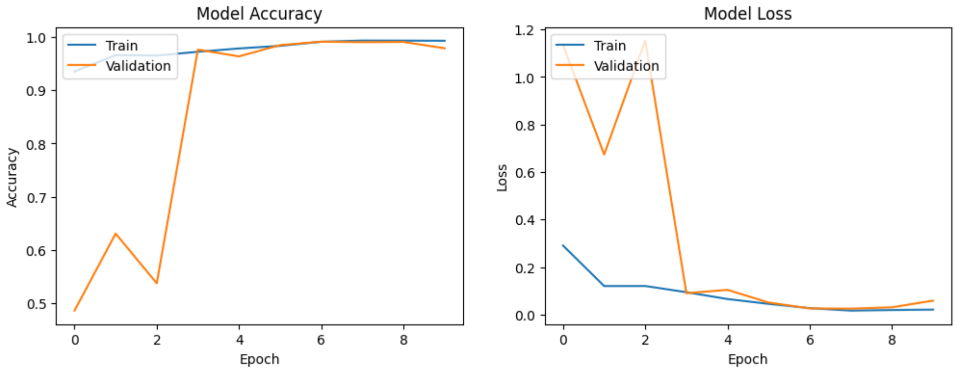

Figure 10: CNN Model Accuracy and loss charts

6.3 Experiment 3 (SOM + CNN)

This experiment is the combination of clustering and classification algorithms and an attempt to replicate the work of Kolesnikov et al. (2024) using the python programming language and minisom library. The idea for these experiments are based on the works of Cheng et al. (2021); Kolesnikov et al. (2024)

Fig 11 shows the histogram of the calculated metrics, which are explained in sections 3.3.

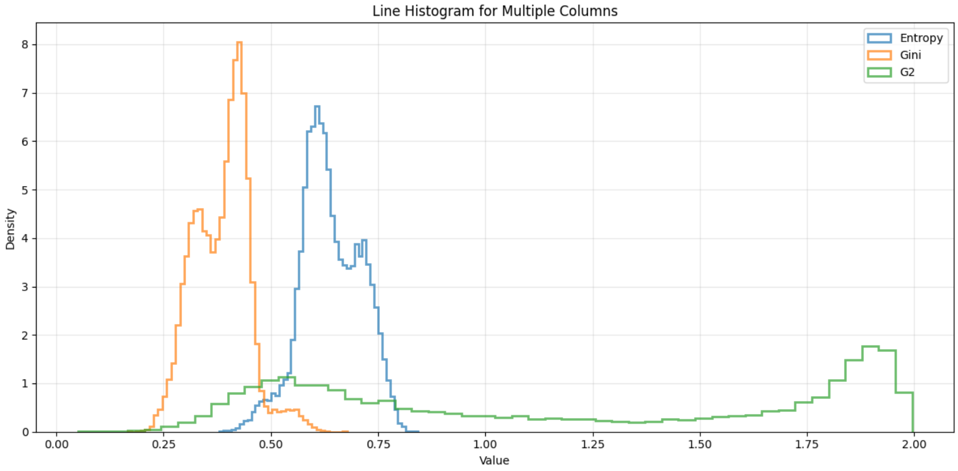

Figure 11: Calculated Metrics: This diagram shows the metric distribution of the galaxies. The skew of **Entropy** and **Gradient pattern** can be seen to the right and the **Gini** to the left, reflecting there being more identified spiral galaxies than elliptical, as explained in 3.3

Fig 12 and 13 shows the distributions of clusters based on the calculated metrics. The clusters having higher gradient pattern metrics have more spiral galaxies in them and they are colored as red in 13. Fig 14 shows the segregation of calculated points for each square of the grid on the som feature map, highlighting high values in red and low in blue.

The dendogram shown in figure15 helps visualise more better the hierarchy and the seperation of clusters.

The results and performance of the cluster algorithm is shown in table 1, by comparing it with the predefined labels.

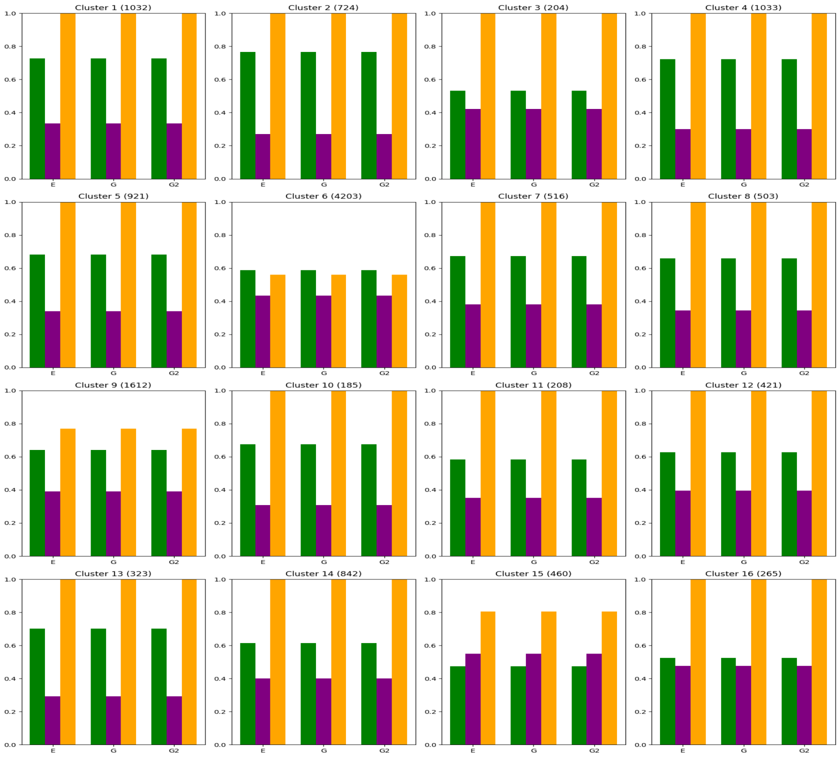

Figure 12: This diagram shows a bar chart of the calculated metrics of the galaxies segregated in different clusters, where **Entropy** is **green**, **gini** is **purple** and **gradient** in **orange**. Charts showing higher gradient are more likely to be spiral clusters and vice versa.

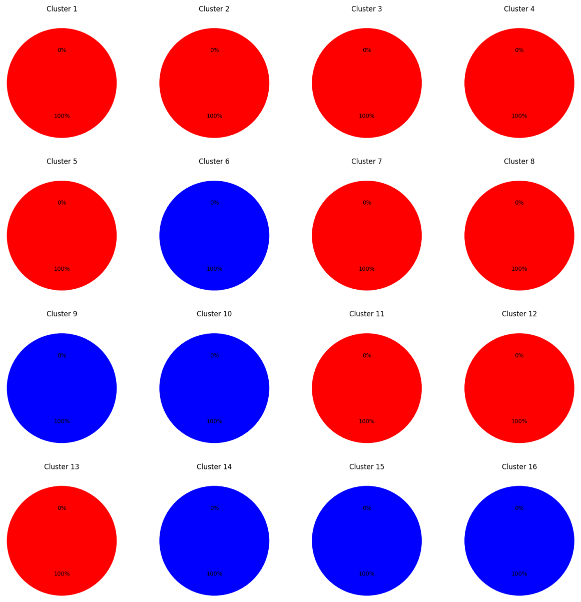

Figure 13: Cluster Distribution plot, red circles signifying clusters with spirals and blue signifying elliptical clusters.

The confusion matrix and model accuracy and loss plots are shown in figures 16 and 17 respectively.

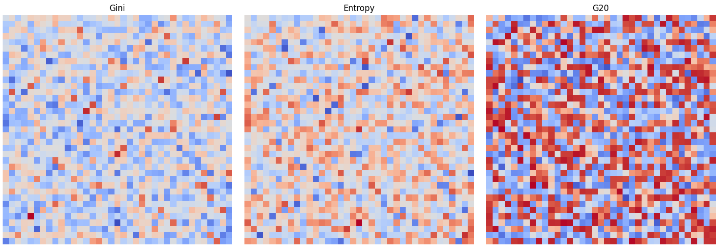

Figure 14: This diagram shows the plotting of cluster points on the SOM feature Map for a cluster grid calculated by the formula $grid\_size = \sqrt{(n\_subjects * 0.1)}$

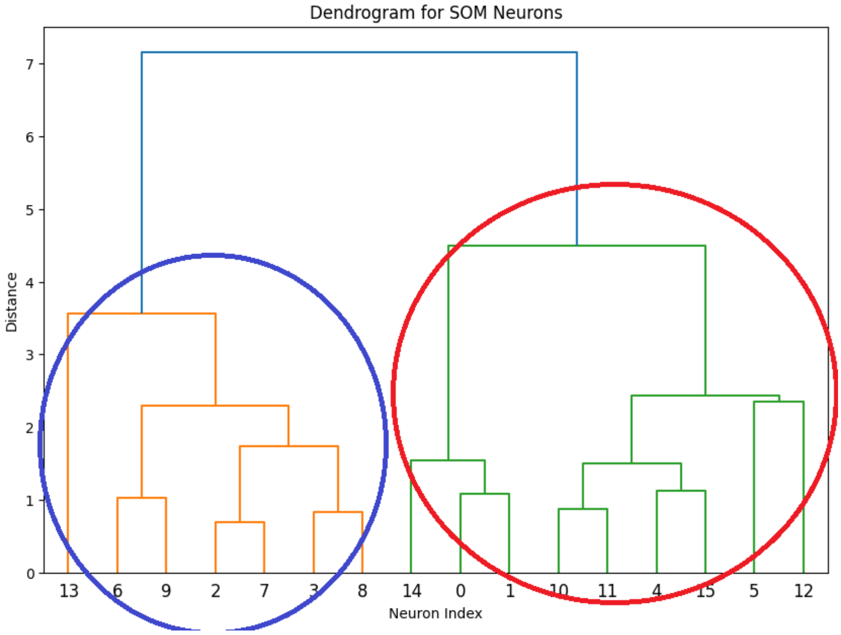

Figure 15: SOM Dendogram: This shows the segregation of cluster on a higher hierarchy level, blue highlights the elipticals and red the spirals

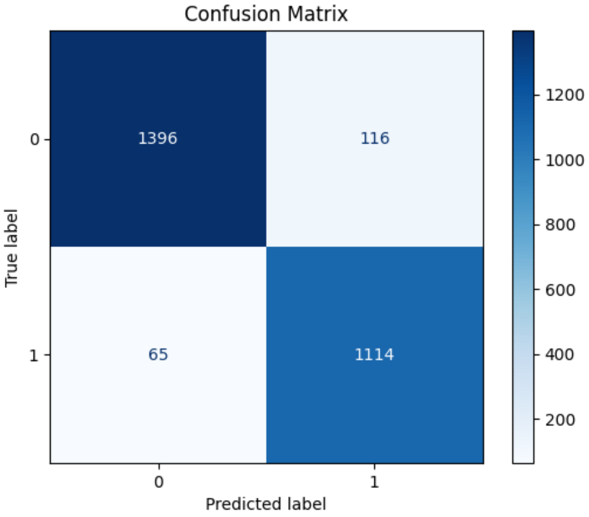

Figure 16: SOM+CNN Confusion Matrix

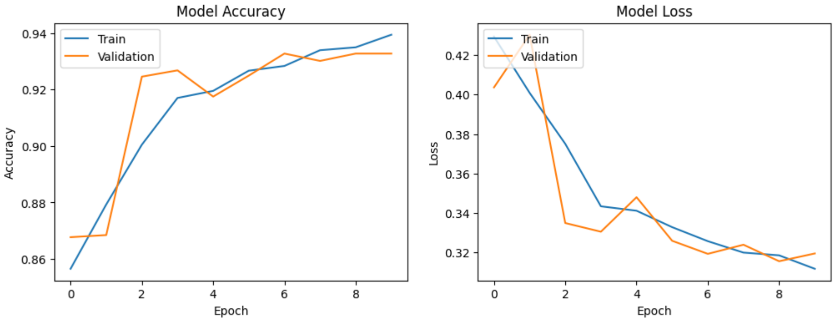

Figure 17: SOM_CNN: Model Loss vs Model Accuracy plot

6.4 Experiment 4 (HDBSCAN + EfficientNetB0)

This experiment uses the HDBScan and efficientNetB0 architectures as the clustering and Classification algorithms. The motivation for using these two algorithms comes from the works of Kalvankar et al. (2021); Tan and Le (2020); Nazeri (2023).

The metrics calculations step is the same as in the above experiment and the results can be seen in Figure 11.

The dendogram for the hdbscan algorithm is shown in fig 18 for $\lambda = 5.0$ as discussed in section 4.2.2

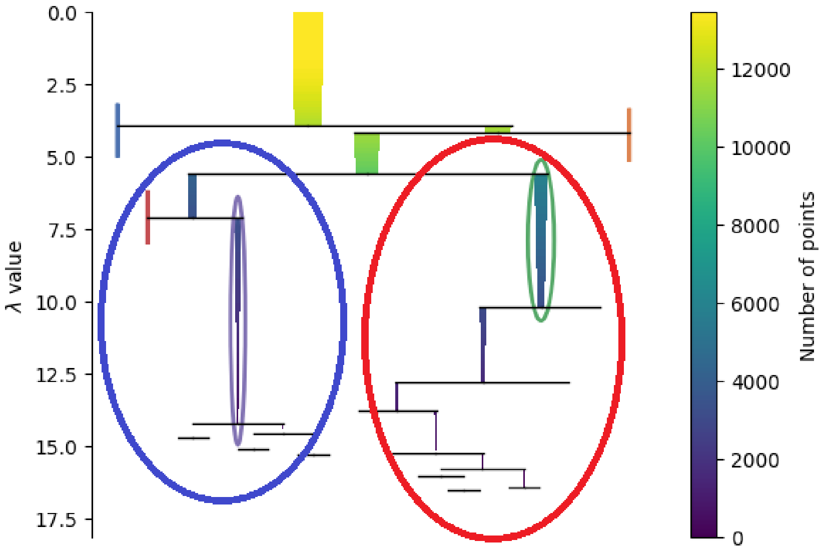

Figure 18: HDBScan Dendogram: This shows the clustered tree plot of hdbscan, the major two clusters are taken from point $\lambda = 5.0$ to keep a moderate distance between data points

Figure 19 and 20 show the confusion matrix and model loss and accuracy charts respectively.

6.5 Discussion

This section shall help understand all the above results in a more detailed way, by comparing the results of all the experiments in a tabular format in tables 1 and 2, out

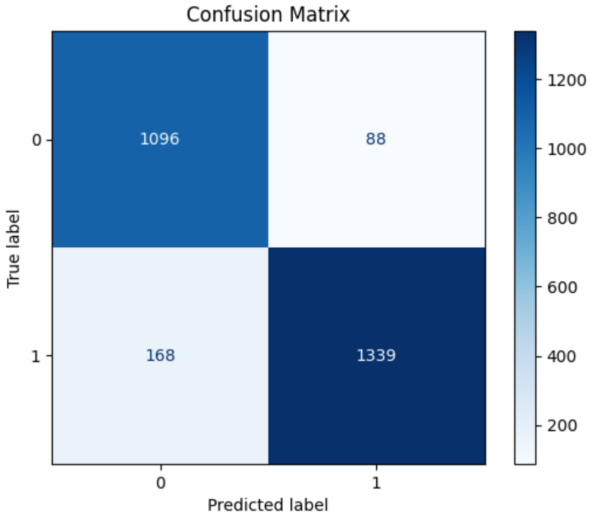

Figure 19: Confusion Matrix of HDBScan + EfficientNetB0

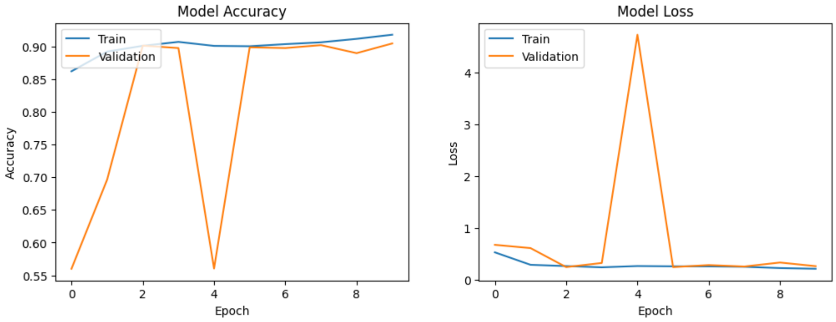

Figure 20: Model loss and accuracy of HDBScan + EfficientNetB0

of which table 1 shows the clustering performance, while table 2 show the classification performances.

The first two experiments were carried out to establish a baseline for the classification algorithms used in the next two experiments, which are the combined methods. The comparison of their performances are shown in table 2.

Algorithm	**Elliptical**	**Spiral**	**True %**
SOM	5885	7567	86.33%
HDBScan	5937	7515	88.38%

Table 1: Clustering Performance Comparison: This shows the galaxies segregated in each cluster and the percentage of rightly labeled galaxies

In table 2 it can be seen that the stand alone classification algorithms perform better than the combined methods, this may be attributed to some amount of mislabelling occuring during the clustering process, as evident in table 1 that show that both clustering algorithms are able to provide accuracy in labelling the galaxies of upto 90%, this signifies that approximately 10% are mislabelled, which thereby brings about complications in training the model which can be seen in figures 17 and 20, which show sudden spikes and drops in model accuracy and loss.

Another notable point is that combined methods still give good results, with an

Algorithm	**Precision**	**Recall**	**F1_Score**	**Accuracy**
CNN	0.98	0.99	0.98	98.81%
EfficientNetB0	0.99	0.96	0.97	97.88%
SOM+CNN	0.92	0.95	0.93	93.27%
HDBScan+EfficientNetB0	0.93	0.87	0.89	90.49%

Table 2: Classificiation Performance Comparison

accuracy close to 90% in both scenarios, as seen in Table 2, which may be an indication that with more research on this ideology, it can be possible to eliminate human biases in the morphological classification of galaxies.

7 Conclusion and Future Work

In conclusion, the objective of the paper to develop a framework for morphological classifications of galaxies to eliminate human biases in the process by using a combination of clustering and classification techniques, has been sufficiently achieved in the paper. Although the paper gives good results, there is still room for improvisations that can be made in the algorithms. Both experiments involving the combined method give a significant accuracy of more than 90%, thereby showing positive signs of this methodology being used for future data surveys such as Large Synoptic Survey Telescope (LSST) [7], Hubble Space Telescope (HST) [8], etc.

This research has focused only on two major classes of galaxies, namely the elliptical and spiral classes; for future work, with the help of domain experts or by understanding other metric distributions, this methodology can be used to recognize and segregate data into more finer galactic classes, like round elliptical, cigar-shaped elliptical, barred spiral, irregulars, etc. Other Clustering algorithms such as Spectral clustering, agglomerative clustering, Gaussian Mixture clustering, etc can be used along with other structures of CNN to analyse the results and try and improve performance.

References

Abraham, R. G., Bergh, S. v. d. and Nair, P. (2003). A New Approach to Galaxy Morphology. I. Analysis of the Sloan Digital Sky Survey Early Data Release - IOPscience. Publisher: IOP Publishing.
URL: *https://iopscience.iop.org/article/10.1086/373919*

Baumstark, M. and Vinci, G. (2024). Spiral-Elliptical automated galaxy morphology classification from telescope images., *Astronomy & Computing* **46**: N.PAG–N.PAG. Publisher: Elsevier B.V.

Becker, B., Vaccari, M., Prescott, M. and Grobler, T. (2021). CNN architecture comparison for radio galaxy classification, *Monthly Notices of the Royal Astronomical Society* **503**(2): 1828–1846.
URL: *https://doi.org/10.1093/mnras/stab325*

[7] LSST: `https://www.lsst.org/`

[8] HST: `https://hubblesite.org/home`

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag, Berlin, Heidelberg.

Chen, C. (2023). Galaxy morphology classification using VGG16, *Journal of Physics: Conference Series* **2580**(1): 012023. Publisher: IOP Publishing.
URL: *https://dx.doi.org/10.1088/1742-6596/2580/1/012023*

Cheng, T.-Y., Huertas-Company, M., Conselice, C. J., Aragón-Salamanca, A., Robertson, B. E. and Ramachandra, N. (2021). Beyond the Hubble Sequence – Exploring Galaxy Morphology with Unsupervised Machine Learning, *Monthly Notices of the Royal Astronomical Society* **503**(3): 4446–4465. arXiv:2009.11932 [astro-ph].
URL: *http://arxiv.org/abs/2009.11932*

Conselice, C. J. (2003). The Relationship between Stellar Light Distributions of Galaxies and Their Formation Histories, *The Astrophysical Journal Supplement Series* **147**: 1–28. Publisher: IOP ADS Bibcode: 2003ApJS..147....1C.
URL: *https://ui.adsabs.harvard.edu/abs/2003ApJS..147....1C*

Dai, Y., Xu, J., Song, J., Fang, G., Zhou, C., Ba, S., Gu, Y., Lin, Z. and Kong, X. (2023). The Classification of Galaxy Morphology in H-band of COSMOS-DASH Field: a combination-based machine learning clustering model. arXiv:2307.02335 [astro-ph].
URL: *http://arxiv.org/abs/2307.02335*

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D. and Fischer, J. L. (2018). Improving galaxy morphologies for SDSS with Deep Learning, *Monthly Notices of the Royal Astronomical Society* **476**(3): 3661–3676.
URL: *https://doi.org/10.1093/mnras/sty338*

Ferrari, F., Carvalho, R. R. d. and Trevisan, M. (2015). MORFOMETRYKA—A NEW WAY OF ESTABLISHING MORPHOLOGICAL CLASSIFICATION OF GALAXIES, *The Astrophysical Journal* **814**(1): 55. Publisher: The American Astronomical Society.
URL: *https://dx.doi.org/10.1088/0004-637X/814/1/55*

Fielding, E., Nyirenda, C. N. and Vaccari, M. (2021). A Comparison of Deep Learning Architectures for Optical Galaxy Morphology Classification, *2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pp. 1–5.
URL: *https://ieeexplore.ieee.org/document/9698414?arnumber=9698414*

Fraix-Burnet, D. (2023). Machine Learning and galaxy morphology: for what purpose?, *Monthly Notices of the Royal Astronomical Society* **523**(3): 3974–3990. arXiv:2306.02626 [astro-ph].
URL: *http://arxiv.org/abs/2306.02626*

Gupta, R., Srijith, P. K. and Desai, S. (2022). Galaxy morphology classification using neural ordinary differential equations, *Astronomy and Computing* **38**: 100543.
URL: *https://www.sciencedirect.com/science/article/pii/S2213133721000858*

Kalvankar, S., Pandit, H. and Parwate, P. (2021). Galaxy Morphology Classification using EfficientNet Architectures. arXiv:2008.13611 [astro-ph].
URL: *http://arxiv.org/abs/2008.13611*

Kolesnikov, I., Sampaio, V. M., de Carvalho, R. R., Conselice, C., Rembold, S. B., Mendes, C. L. and Rosa, R. R. (2024). Unveiling Galaxy Morphology through an Unsupervised-Supervised Hybrid Approach, *Monthly Notices of the Royal Astronomical Society* **528**(1): 82–107. arXiv:2401.08906 [astro-ph].
URL: *http://arxiv.org/abs/2401.08906*

Lahav, O., Naim, A., Buta, R. J., Corwin, H. G., de Vaucouleurs, G., Dressler, A., Huchra, J. P., van den Bergh, S., Raychaudhury, S., Sodré, L. and Storrie-Lombardi, M. C. (1995). Galaxies, Human Eyes, and Artificial Neural Networks, *Science* **267**(5199): 859–862. Publisher: American Association for the Advancement of Science.
URL: *https://www.science.org/doi/10.1126/science.267.5199.859*

Lotz, J. M., Primack, J. and Madau, P. (2004). A New Nonparametric Approach to Galaxy Morphological Classification - IOPscience. Publisher: IOP Publishing.
URL: *https://iopscience.iop.org/article/10.1086/421849*

Lukic, V., Brüggen, M., Banfield, J. K., Wong, O. I., Rudnick, L., Norris, R. P. and Simmons, B. (2018). Radio Galaxy Zoo: compact and extended radio source classification with deep learning, *Monthly Notices of the Royal Astronomical Society* **476**(1): 246–260.
URL: *https://doi.org/10.1093/mnras/sty163*

Ma, X., Li, X., Luo, A., Zhang, J. and Li, H. (2023). Galaxy image classification using hierarchical data learning with weighted sampling and label smoothing, *Monthly Notices of the Royal Astronomical Society* **519**(3): 4765–4779.
URL: *https://academic.oup.com/mnras/article/519/3/4765/6957257*

Nazeri, S. (2023). Comparing The-State-of-The-Art Clustering Algorithms.
URL: *https://medium.com/@sina.nazeri/comparing-the-state-of-the-art-clustering-algorithms-1e65a08157a1*

Patel, J. (2023). *Classifying Galaxy Images Using Improved Residual Networks*, PhD thesis, University of Windsor.
URL: *https://scholar.uwindsor.ca/etd/9112*

Rosa, R. R., de Carvalho, R. R., Sautter, R. A., Barchi, P. H., Stalder, D. H., Moura, T. C., Rembold, S. B., Morell, D. R. F. and Ferreira, N. C. (2018). Gradient pattern analysis applied to galaxy morphology, *Monthly Notices of the Royal Astronomical Society: Letters* **477**(1): L101–L105.
URL: *https://doi.org/10.1093/mnrasl/sly054*

Rosito, M. S., Bignone, L. A., Tissera, P. B. and Pedrosa, S. E. (2023). Application of dimensionality reduction and clustering algorithms for the classification of kinematic morphologies of galaxies, *Astronomy & Astrophysics* **671**: A19.
URL: *https://www.aanda.org/10.1051/0004-6361/202244707*

Sautter, R. A. (2018). GRADIENT PATTERN ANALYSIS: NEW METHODOLOGICAL AND COMPUTATIONAL FEATURES AND APPLICATION.

Savić, V., Jankov, I., Yu, W., Petrecca, V., Temple, M. J., Ni, Q., Shirley, R., Kovacevic, A. B., Nikolic, M., Ilic, D., Popovic, L. C., Paolillo, M., Panda, S., Ciprijanovic, A. and Richards, G. T. (2023). The LSST AGN Data Challenge: Selection methods. arXiv:2307.04072 [astro-ph].
URL: *http://arxiv.org/abs/2307.04072*

Shi, J.-H., Qiu, B., Luo, A.-L., He, Z.-D., Kong, X. and Jiang, X. (2023). Stellar classification with convolutional neural networks and photometric images: a new catalogue of 50 million SDSS stars without spectra, *Monthly Notices of the Royal Astronomical Society* **520**(2): 2269–2280.
URL: *https://academic.oup.com/mnras/article/520/2/2269/7000842*

Tan, M. and Le, Q. V. (2020). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946 [cs, stat].
URL: *http://arxiv.org/abs/1905.11946*

Tang, H., Scaife, A. M. M., Wong, O. I. and Shabala, S. S. (2022). Radio Galaxy Zoo: giant radio galaxy classification using multidomain deep learning, *Monthly Notices of the Royal Astronomical Society* **510**(3): 4504–4524.
URL: *https://doi.org/10.1093/mnras/stab3553*

Wu, D., Zhang, J., Li, X. and Li, H. (2022). A Lightweight Deep Learning Framework for Galaxy Morphology Classification, *Research in Astronomy and Astrophysics* **22**(11): 115011. Publisher: National Astromonical Observatories, CAS and IOP Publishing.
URL: *https://dx.doi.org/10.1088/1674-4527/ac92f7*

Yaa, Noirot Céline, H. J. M. J.-A. K. M. , Fanny, Guilmineau Camille, K. A. M. D. S. and Nathalie, Gaspin Christine, L. L. V. H.-S. (2023). *Section 11 Self-Organizing Map (SOM) | ASTERICS: User documentation.*
URL: $https://asterics.pages.mia.inra.fr/user_documentation/som.htmlref - vialaneixWSOM2017$

Yu, H. and Hou, X. (2022). Hierarchical clustering in astronomy, *Astronomy and Computing* **41**: 100662.
URL: *https://www.sciencedirect.com/science/article/pii/S2213133722000762*

Zhu, X.-P., Dai, J.-M., Bian, C.-J., Chen, Y., Chen, S. and Hu, C. (2019). Galaxy morphology classification with deep convolutional neural networks., *Astrophysics & Space Science* **364**(4): N.PAG–N.PAG. Publisher: Springer Nature.