Galaxy Morphology Classifier Proposal

David Thornton

B00152842

November 2025

0.1 Declaration

I/We declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, except where otherwise stated. I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. I/We understand that plagiarism, collusion, and copying are grave and serious offences and accept the penalties that would be imposed should I/we engage in plagiarism, collusion or copying. I acknowledge that copying someone else's assignment, or part of it, is wrong, and that submitting identical work to others constitutes a form of plagiarism. I/We have read and understood the colleges plagiarism policy 3AS08 (available here). This material, or any part of it, has not been previously submitted for assessment for an academic purpose at this or any other academic institution. I have not allowed anyone to copy my work with the intention of passing it off as their own work.

Name: David Thornton Dated: 20/11/2025

Contents

1 Introduction

Galaxies can be classified in many different ways with different techniques. One of the most common ways is to classify them by their morphology, their physical components and structure Conselice 2014. Classifying galaxies this way can provide insight into their formation and life cycle, as well as the general composition of the universe Conselice 2014. However, there are a massive amount of galaxies in our sky, many of which have already been captured as images Tegmark et al. 2004. The time and resources it would take for these images to be classified by people would be truly immense. A better alternative to this is to train a machine learning (ML) model to carry out this task Serrano-Pérez, Díaz Hernández, and Sucar 2024. However, ML models are not as trustworthy as people, and for scientific applications such as research of galaxy systems this can be an issue Marusich et al. 2024. The aim of this literature review is to evaluate modern techniques of galaxy image classification, including image feature extraction, hierarchical classification, uncertainty estimation and ML explainability.

High quality training data is an extremely important factor in model accuracy within and ML task, including image classification Theng and Bhoyar 2024. This literature review will discuss modern feature extraction techniques in images, including a discussion of common features identified in galaxies, along with an analysis of image selection, dimensionality reduction, manual and automatic feature extraction techniques in the current literature.

Morphological methods of galaxy classification often follow a hierarchical structure, which can be embedded into ML models Aguilar-Argüello et al. 2025. Doing so not only leads to a better understanding of model decision but can also lead to enhanced model performance, although coming at the cost of larger and hence more computationally demanding models Serrano-Pérez, Díaz Hernández, and Sucar 2024. This literature review will discuss modern implementation strategies of hierarchical ML models, including a discussion of hierarchies in galaxy morphology classifications, recent implementations of hierarchical structuring within galaxy classification models and the benefits and drawbacks of both different implementation methods of hierarchical models and how they compare to flat models.

Most classification systems, including galaxy classification systems, simply output a single predicted class Tyralis and Papacharalampous 2024. This limited information can lead to misinterpretations of model predictions, particularly when class predictions are borderline or uncertain, and generally does not provide sufficient information to gain a full understanding of the model's predictions Marusich et al. 2024. The implementation of uncertainty quantification can rectify this issue, by providing more insight into the model's confidence in its predictions Marusich et al. 2024. This literature review will discuss uncertainty estimation techniques in recent galaxy classification studies, including an overview of the techniques used and a comparison of their performance relative to each other and systems without uncertainty estimation.

Image classification systems often use neural networks for classification and

feature extraction since they often perform well at understanding more complex data such as images Yin et al. 2024. However, an issue with these models is that, due to their complexity, they are inherently uninterpretable, meaning how they come to their predictions is unknown Retzlaff et al. 2024. This uninterpretability means that the model may not be identifying the correct traits to distinguish between images and will not provide any insight into how decisions are made Yin et al. 2024. To rectify this issue, explainability techniques, or inherently interpretable model must be used Cao et al. 2025. This literature review will discuss current explainability techniques used in galaxy classification systems and directly compare these techniques.

2 Image Feature Extraction

One of the most crucial components of any ML task is to have meaningful features for the model to use in its prediction Theng and Bhoyar 2024. In more basic ML models, this involves understanding the available features and selecting the most meaningful or manipulating existing features to make better ones Theng and Bhoyar 2024. However, in image classification systems such as the ones examined in this literature review, there are no features available aside from the images themselves. Because of this, usable features must be extracted from these images Yin et al. 2024. The following section will outline the techniques used in current literature in this process.

2.1 Common features of galaxy images

To gain an understanding of the types of features that should be extracted from the image, the features or traits humans look for when classifying galaxy images can be used as a starting point, such as those in datasets used in ML training.

Some common features looked at in galaxy image catalogues, such as the EFIGI Catalogue Baillard et al. 2011 or Nair and Abraham Catalogue Nair and Abraham 2010 are as follows:

- Inclination/elongation
- Environment
- Bulge
- Spiral Arm Properties
- Luminosity/Texture

The inclination is a property in disk shaped galaxies that describes the tilt of the galaxy relative to us, with a high inclination meaning the galaxy is closer to being on edge. The elongation is a property in non-disk-shaped galaxies that describes the ratio between the major and minor axis, where a higher elongation means the galaxy has a high major to minor axis ratio Baillard et

al. 2011. This feature is most important when classifying the sub classifications of elliptical galaxies and is also useful in identifying if a disk-shaped galaxy is on edge or not.

The environment is a description of whether there are any obstructing/interfering objects in the image, such as other stars or galaxies Baillard et al. 2011. This feature is not as relevant in classification but recognizing whether an image is obstructed or not is highly important in data quality.

The bulge is a bright cluster of stars in the center of a galaxy. Features of the bulge include its size relative to the rest of the galaxy and its shape (bar or round) Nair and Abraham 2010. This feature is most important in classifying spiral galaxy subtypes and in classifying galaxies as elliptical or other (since elliptical galaxies have large bulges).

Spiral arm properties can be broken down into a few subtypes:

- Arm strength: how bright the arms appear relative to the rest of the galaxy
- Arm curvature: how tightly wound the arms are
- Rotation: whether arms rotate clockwise or anticlockwise

Conselice 2014
These features are most important in classifying galaxies as spiral or other, the sub categorization of spiral galaxy.

Luminosity and texture describe the amount of light coming from the galaxy and the structure/formation of that light. This can include:

- Overall Luminosity: how bright the galaxy is
- Hot spots: brighter sections/clumps in the galaxy
- Dispersion of light: How uniform the dispersion of light is

Some of these features are related to others previously mentioned Nair and Abraham 2010. These features can be used to identify the main galaxy type and sub types within spiral galaxies.

All these feature/traits should be kept in mind during the feature extraction/selection process.

2.2 Image Selection and Dimensionality Reduction

Image selection and dimensionality reduction are important steps in creating features for computer vision systems Ali, Wassif, and Bayomi 2024. Choosing images that are good representations of classes to be labelled will increase the accuracy of the model by providing data of a sufficient quality Feuer et al. 2024. Dimensionality reduction, the process of reducing the dimensionality, or size, of images is also an important step in feature creation as it removes noise from the data and reduces the amount of processing the model must do during training

and predicting Ali, Wassif, and Bayomi 2024. There are different approaches that can be taken in image selection and dimensionality reduction.

Some datasets already undergo specific image selection. In the catalogue created by Nair and Abraham Nair and Abraham 2010 only images with a brightness of 16 mag on the g-band magnitude scale or brighter were used and they also removed any images with a redshift values greater than 0.1. The EFIGI catalogue Baillard et al. 2011 only used images with five bands of light captured, with the three visible bands (red, green and blue) along with ultraviolet and infrared light. Both also removed all images they classified as non-galaxy images. Some studies also included their own image selection process on already defined datasets, such as by removing underrepresented classes and removing rows from overrepresented classes Huang, Wu, and Huang 2024.

There was a much greater emphasis in dimensionality reduction in the reviewed literature then image selection, with many different techniques being implemented. Many papers reduced the dimensionality removing pixels that did not represent the galaxy. Some did this by only selecting the central pixels from the image, with one paper finding the 95 percent confidence interval for the mean value of the parameter was estimated as 53 ± 2 pixels, corresponding to a window size of 106 ± 4 pixels Semenov et al. 2025. Dimensionality reduction through grey scaling, converting pixel values into a single dimension for colour by combining the colour values across the RGB scale Semenov et al. 2025 and through normalizing the pixel values using the min max algorithm Dias 2025. Another paper described isolating the galaxy pixels and removing all others using Adaptive Selective Min Filter, which involves first removing any pixels which are not in bright neighbourhoods of pixels, then the galaxy in the image is identified by selecting the largest bright area of pixels near the center of the image, and pixels outside of this area, or that brighter or dimmer compared to it are removed Sarkar, Palit, and Bhattacharya 2024. Another paper compared the effectiveness of several techniques, including Local Linear Embedding (LLE), Isomap, and Principal Component Analysis (PCA) which produced similar results as well as Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbour Embedding (t-SNE) Semenov et al. 2025.

2.3 Methods of feature extraction

The most prominent feature extraction method across literature in this review was automated techniques, which involve training the model to learn features from the images themselves. The most frequently used type of model was Convolutional Neural Networks (CNNs) Serrano-Pérez, Díaz Hernández, and Sucar 2024, Huang, Wu, and Huang 2024, Mohan and Scaife 2024, Goh et al. 2025, Sarkar, Palit, and Bhattacharya 2024, Semenov et al. 2025, Atemkeng et al. 2025 and Feuer et al. 2024, which are often used to extract features from images. Some studies used pre trained CNNs, or CNNs that have already been trained on large datasets and hence have already gained the ability to detect features like shapes and edges Serrano-Pérez, Díaz Hernández, and Sucar 2024.

Different types of CNNs were also used, such as C2 type model Sarkar, Palit, and Bhattacharya 2024, EfficientNet Serrano-Pérez, Díaz Hernández, and Sucar 2024 and LeNet-5 with an additional layer added Mohan and Scaife 2024.

However, other feature extraction techniques were used, both before being passed into an automated feature extraction model and traditional models. Clustering was used in two papers, allowing the classification model to learn to associate different clusters with different classes Marín Díaz, Gómez Medina, and Aijón Jiménez 2024. One such paper tested both Self-Organizing Maps and Hierarchical Density Based Scan Mohan and Scaife 2024. Some papers also outlined manual feature extraction techniques, such as an asymmetry feature, which was calculated by comparing the absolute pixel difference between an image and itself rotated 180 degrees. This was used to identify irregular galaxies Sarkar, Palit, and Bhattacharya 2024. Another manually extracted feature from the same paper was a tidal feature, which subtracted a Gaussian blur of the image from the greyscale of the same image to reveal the prominent features, which was useful in identifying spiral subclasses Sarkar, Palit, and Bhattacharya 2024. Another paper used algorithms to calculate the entropy coefficient, a metric that represents the randomness of flux distributions, Gini coefficient, the distribution of light across the image, and gradient pattern which measures the asymmetry of the image by comparing pixels that are equidistant from the center Dias 2025. The Gini coefficient, gradient pattern (also called asymmetry gradient) and asymmetry are also calculated in another paper Aguilar-Argüello et al. 2025, along with other metrics for colour gradient, which describes how the colour of a galaxy differs from the center to the outside, specifically comparing light in the infrared and ultraviolet wavelengths and colour index, which looks at colour of the galaxy generally across the whole spectrum, both of which indicate the age of the galaxy. The paper also calculated concentration, how centrally concentrated light is and clumpiness, or the fraction of light within clumpy regions, which can be useful in identifying spiral galaxies Aguilar-Argüello et al. 2025.

2.4 Comparison of performance with different features/extraction techniques

The results from these studies have indicated that the resulting model performance is higher when preprocessing and manual feature extraction steps are implemented alongside automated feature extraction using models Sarkar, Palit, and Bhattacharya 2024. This is most clearly indicated in papers which gave direct comparisons of model performance with and without preprocessing and feature extraction. One such paper found a direct correlation between the number of features extracted using different extraction techniques and the overall model performance, with t-SNE having 54 percent accuracy with 2 components, and LLE having 93 percent accuracy with 138 components Semenov et al. 2025.

Accuracy	**Method**	**Components**	**Neighbours**	**Perplexity**
0.93	LLE	138	10	N/A
0.92	UMAP	82	26	N/A
0.88	Isomap	87	10	N/A
0.61	PCA	63	N/A	N/A
0.54	t-SNE	2	102	33

Table 1: Fine-tuned parameters for each of the dimensionality reduction methods used. The classification method in the pipeline is logistic regression, the classification case is round/in-between/cigar, Semenov et al. 2025 Page 3

Another paper directly compared performance with and without preprocessing and found up to 10 percent accuracy increase when classifying 10 or 4 classes, in some models, with preprocessing, as opposed to without it Sarkar, Palit, and Bhattacharya 2024.

Model	**Type of input image**	**Accuracy**	
		4 classes	10 classes
C2 model	Pre-processed	74.40%	59.48%
	Original	70.92%	48.38%
Modified LeNet-5	Pre-processed	80.31%	68.31%
	Original	76.31%	59.56%
CoAtNet-0	Pre-processed	90.33%	92.38%
	Original	83.36%	80.40%
CoAtNet-0-R	Pre-processed	93.13%	94.31%
	Original	83.78%	78.08%

Table 2: Accuracies of different neural networks on Pre-processed and Original images Sarkar, Palit, and Bhattacharya 2024 Page 4

The same paper also found that the more manually extracted features where used, the better the performance, with or without preprocessing Sarkar, Palit, and Bhattacharya 2024.

Model	Input Type	Feature	10 Class Accuracies (overall)
C2	Pre-processed	F1	64.86%
		F2	63.02%
		F1 & F2	65.29%
	Original	F1	50.23%
		F2	47.79%
		F1 & F2	57.38%
Modified LeNet-5	Pre-processed	F1	69.62%
		F2	65.63%
		F1 & F2	70.23%
	Original	F1	58.94%
		F2	57.50%
		F1 & F2	60.17%
CoAtNet-0	Pre-processed	F1	90.69%
		F2	89.65%
		F1 & F2	93.36%
	Original	F1	85.97%
		F2	79.12%
		F1 & F2	89.69%
CoAtNet-0-R (Proposed)	Pre-processed	F1	90.00%
		F2	91.72%
		F1 & F2	92.93%
	Original	F1	86.98%
		F2	79.6%
		F1 & F2	89.68%

Table 3: Accuracies of Neural Networks models stacked with Features Sarkar, Palit, and Bhattacharya 2024 Page 4

Other papers also showed a similar trend, with models which used feature extraction having higher accuracy than ones which did not. For example, one paper which used manual feature extraction techniques had a F1 score of 98 percent Dias 2025 while another that did not had a score of 83 percent Serrano-Pérez, Díaz Hernández, and Sucar 2024, even though both used the CNN architecture.

It should also be noted however, that these preprocessing and feature extraction steps are resource intensive, especially when done with the number of images used when training image classification models Yin et al. 2024.

This section of the literature review has found that most galaxy classification models used some kind of automated feature extraction using models such as CNNs. However, models which used a combination of automated and manual feature extraction performed better than those with only one feature extraction method. This is clearly indicated both in models which compared the performance with only one feature extraction method with several Sarkar, Palit, and Bhattacharya 2024 and in models which only performed manual Aguilar-Argüello et al. 2025 or automatic Huang, Wu, and Huang 2024 feature extraction underperforming compared to other models. This fully automatic feature

extraction model also removed underrepresented classes from the training data. Although this may have increased model accuracy, this is not ideal in a classification task where these underrepresented classes must still be identified Huang, Wu, and Huang 2024. Many of these models which relied heavily on automatic feature extraction techniques also lacked interpretability techniques, which is an issue considering the lack of interpretability in the black box models used Serrano-Pérez, Díaz Hernández, and Sucar 2024. Studies which used smaller models also had a tendency to underperform on underrepresented classes, likely due to underfitting as the model could not learn patterns from all classes Goh et al. 2025. Despite this, other models likely could have benefited from image selection techniques, particularly in datasets labelled by the public where removing rows with less certain classifications, where different classes had a similar number of votes Sarkar, Palit, and Bhattacharya 2024. Finally, some studies lacked any preprocessing/dimensionality reduction techniques, which likely would have led to increased performance Mohan and Scaife 2024.

3 Hierarchical ML Models

Many classification tasks, such as galaxy classification, follow a hierarchical structure where classes can be divided into broad categories and subcategories within these higher level categories. This structure can also be represented in ML models and is often used in more basic classification models Miranda, Köhnecke, and Renard 2023. However, this structure can also be implemented in image classification tasks like galaxy classification Serrano-Pérez, Díaz Hernández, and Sucar 2024. The following section outlines modern techniques to implement this with galaxy morphology classification tasks. Even though galaxy classifications have a hierarchical structure, most models used in predicting galaxy classes have a flat architecture, likely because flat models are simpler to implement and are more common in image classification systems Aguilar-Argüello et al. 2025. In this review two hierarchical models will be analysed. The first is based on a Bayesian network combined with a CNN for feature extraction Serrano-Pérez, Díaz Hernández, and Sucar 2024, which will be referred to as the BCNN model. The second uses an eXtreme Gradient Boost (XGBoost) model for classification, with features taken from the Nair and Abraham Catalogue Nair and Abraham 2010 and the Visual Morphology Catalogue for the Mapping Nearby Galaxies Vázquez-Mata et al. 2022. The XGBoost model is not inherently hierarchical but uses several XGBoost models to represent a hierarchy. However, the features used in this model are similar to the features extracted in other papers mentioned in section 2 of this review Aguilar-Argüello et al. 2025. This model will be refer to as the XGBoost model.

3.1 Comparison of Flat and Hierarchical structure in machine learning

Flat architectures in machine learning tasks are more common than hierarchical architectures, likely because they are easier to implement Miranda, Köhnecke, and Renard 2023. Flat ML models classify classes that have no relationship to each other or at least treat related classes as if they have no relation Cao, Feng, and An 2024. Hierarchical models on the other hand classify classes according to a hierarchical structure, where all labelled classes belong to a parent class, and all child classes have similarities to the other children of the same parent Miranda, Köhnecke, and Renard 2023. Hierarchical models are often partially or entirely designed with this same structure, either having a sub model representing each node or a model or sub model dedicated to this hierarchy, such as a Bayesian network Serrano-Pérez, Díaz Hernández, and Sucar 2024. Flat models do not have any hierarchy in their structure, ending in a single multi-class prediction layer Cao, Feng, and An 2024.

3.2 Hierarchy in Galaxy classification

The most prominent galaxy classification system used in nearly all classification models today is the Hubble tuning fork Conselice 2014. Although other classifications are often used (as mentioned in Section 2), they are not generally associated with any larger classification scheme. The Hubble tuning fork has been modified greatly since it was first proposed, but a common representation is given below:

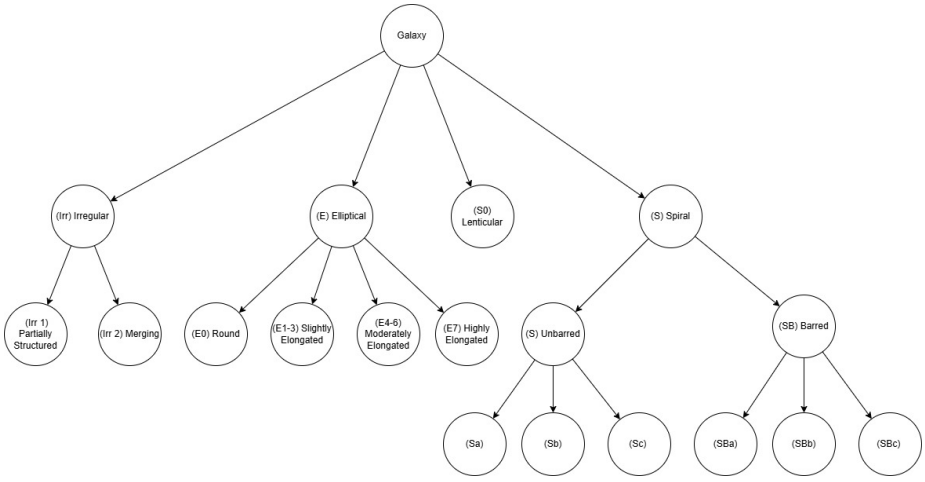

Figure 1: Modern Hubble Sequence

The subcategories are labelled as follows:

- Elliptical subclasses are labelled 0 to 7, where higher subclasses are more elongated.
- Lenticular galaxy subclasses are represented as S0-, S0 and S0+, which represents an increasing level of defined disk structure in the galaxy.
- Spiral subcategories (for both barred and unbarred) are labelled a to c, where the a type has a large bulge and faint, tightly wound arms, the c type has a small bulge and prominent, loose arms and b is in between.

Conselice 2014

3.3 Hierarchical structure in classification systems

The first of the models in this review with a hierarchical structure, the BCNN model. This model considered the following structure in its hierarchy (although it was not explicitly stated whether this was the final model used) Serrano-Pérez, Díaz Hernández, and Sucar 2024.

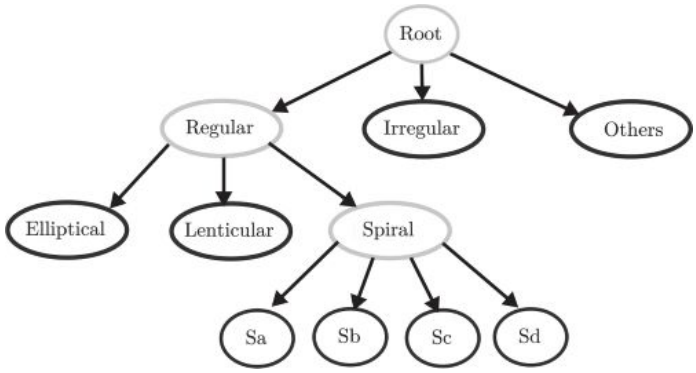

Figure 2: BCNN Hierarchical Structure Serrano-Pérez, Díaz Hernández, and Sucar 2024 Page 10

The second hierarchical model, the XGBoost model, uses the following hierarchical structure Aguilar-Argüello et al. 2025.

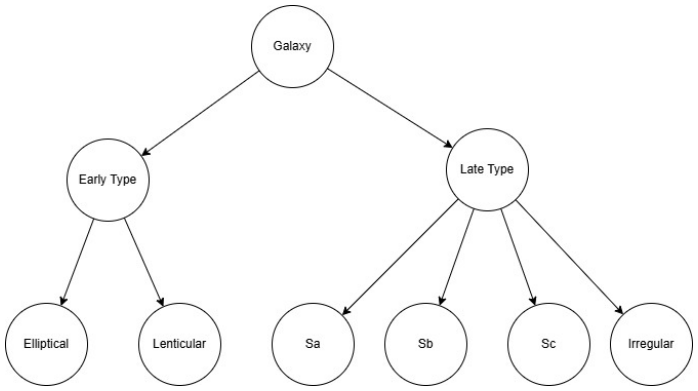

Figure 3: XGBoost Hierarchical Structure Aguilar-Argüello et al. 2025

Both models have similar final subcategories of classes, but the approach taken to get to these leaf classifications differs. The BCNN model used a hierarchy more like the Hubble tuning fork hierarchy, with each layer of the model dividing the classifications into the categories and subcategories in the Hubble tuning fork Serrano-Pérez, Díaz Hernández, and Sucar 2024. The XGBoosted model took a different approach. In this model, the class divisions were instead broken down based on how visually similar categories were, first breaking down into early types, which appear smoother, and later types, which are more structured in appearance. The next layer broke the categories into the subclasses seen in the Hubble tuning fork Aguilar-Argüello et al. 2025.

3.4 Methods of combining hierarchies in hierarchical systems

The approaches taken in implementing the hierarchical structure described above also varied greatly between the two models. The BCNN model used a CNN to predict the subclassifications in the tree, essentially giving the child nodes as a prediction first. The model then combines the likelihood of each child class to get a prediction of the parent class in a bottom-up approach until the leaf node is reached, with a probability of 1. The model then goes back down the tree, going to whichever child node has the highest probability until a leaf node is reached. This leaf node is the final classification chosen Serrano-Pérez, Díaz Hernández, and Sucar 2024. The XGBoosted model takes a different approach. In this model, manually extracted features are passed into an XGBoost model, which performs binary classification on the data, predicting whether the galaxy is an early or late type. The features are then passed into one of two additional XGBoosted models, one classifies early galaxies as either elliptical or lenticular, and the other which classifies the late galaxies into 3 spiral types Aguilar-Argüello et al. 2025.

3.5 Comparison of performance with different structures

Both studies which used hierarchical models compared the model performance with flat models which use the same features/feature extraction. The BCNN model found an increase in model accuracy using the hierarchical approach, with a 7 percent increase in exact matches and a 3 percent increase in the F score Serrano-Pérez, Díaz Hernández, and Sucar 2024. The XGBoost model found no difference or at least minimal difference between the hierarchical and flat models' performance, giving no clear indication of the benefits of this approach Aguilar-Argüello et al. 2025.

Comparing these hierarchical models to other flat models, both compare well with other models which classified similar numbers of galaxies. The BCNN network had a final F score of 81 percent in the hierarchical model, and this score was also achieved with no feature extraction techniques Serrano-Pérez, Díaz Hernández, and Sucar 2024. A similar model, a LeNet-5 model, which had a similar number of label classes had an accuracy of 68 percent, even when feature extraction techniques were used Sarkar, Palit, and Bhattacharya 2024.

Overall, the BCNN showed promising increases in model accuracy when compared to baseline models in the same study and to similar models in other studies Serrano-Pérez, Díaz Hernández, and Sucar 2024. Although the XGBoosted model did not show any meaningful increase in accuracy compared to its baseline models Aguilar-Argüello et al. 2025, similar methods using more in-depth feature extraction techniques may provide better results.

The BCNN model was structured to reflect the Hubble tuning fork exactly, however this structure led to major biases against underrepresented classes, especially at the root node for the irregular and other classes, which only accounted for 9.3 percent and 2.5 percent of the data, with the other 88.2 percent being made up of regular galaxies Serrano-Pérez, Díaz Hernández, and Sucar 2024. To account for this, image augmentation was used. This can be harmful as it may lead to overfitting or inflated performance estimates Szeliski 2022, which may have happened in this study considering initial models did not improve performance with this change Serrano-Pérez, Díaz Hernández, and Sucar 2024. Additionally, although this model was accurate compared to others, there was a severe lack of explainability in the model.

The XGBoost model underperformed compared to the other models, achieving an accuracy of 63 percent Aguilar-Argüello et al. 2025. This model was also the only one which did not use automatic feature extraction, and this coupled with the fact that only 12 features were manually extracted may have led to the model not capturing the full complexity of images Aguilar-Argüello et al. 2025.

4 Uncertainty quantification

Most ML classification systems output a single prediction, returning a single predicted class Tyralis and Papacharalampous 2024. This provides a clear, but

basic interpretation of the classification. In classification tasks, such as galaxy classification, when classes are less distinct Serrano-Pérez, Díaz Hernández, and Sucar 2024, more detailed descriptions of the classification, expressed as confidences in each class, can provide deeper insight when classifications are being assessed especially when confidences are low or borderline, Marusich et al. 2024 since a single prediction will not express this uncertainty. Uncertainty quantification is the process of developing models which will output these confidence intervals. The following section will outline modern uncertainty quantification techniques in galaxy morphology classification tasks.

4.1 Methods used

The most common method of uncertainty quantification found in this literature review was the use of Bayesian uncertainties, which were implemented by Serrano-Pérez, Díaz Hernández, and Sucar 2024 and Zhou et al. 2025. One paper used a CNN to produce probabilistic outputs of predicted classes and then propagated these predictions through a hierarchical galaxy Bayesian network head, combining the probabilistic values of subcategories to calculate the probabilities of parent classes. The model then backpropagated through the tree, following the highest probabilities to find the most likely leaf node Serrano-Pérez, Díaz Hernández, and Sucar 2024. Another paper similarly used a CNN for probabilistic outputs and propagated the results through a Bayesian Network (BN) head, testing two types of BN, Monte Carlo dropout, dropping random nodes in layers of the network during testing and training, simulating different network configurations Tyralis and Papacharalampous 2024, and multiplicative normalizing flows, using normalizing flows to change weight distributions into more complex forms Zhou et al. 2025. Another paper experimented with different Bayesian methods, including Hamiltonian Monte Carlo (HMC), Variational Inference (VI), Last Layer Laplace Approximation (LLA) and Monte Carlo dropout Mohan and Scaife 2024.

The only paper in this review which did not use a Bayesian based approach instead used fuzzy C means clustering Marín Díaz, Gómez Medina, and Aijón Jiménez 2024, although this is an unsupervised approach, not a classification method. C means clustering, similar to K means clustering, categorizes data into model defined clusters, where C means clustering, unlike K means clustering, allows data to fall into several clusters, with different membership values for each Marín Díaz, Gómez Medina, and Aijón Jiménez 2024. This model used this cluster information as inputs to a random forest classification model.

4.2 Comparison of performance with uncertainty techniques

The models which implemented uncertainty quantification techniques in this review generally found success. The first paper discussed, Serrano-Pérez, Díaz Hernández, and Sucar 2024 found that the addition of the BN and probabilistic inference improved the model performance (from its base CNN) by 10 percent

on the exact match scores and 7 percent on F scores. Another paper which compared the performance of several Bayesian techniques Mohan and Scaife 2024 found that most of these techniques performed better than the non-Bayesian baseline, which did not use any uncertainty estimation techniques.

Inference	**Error (%)**	**UCE**
HMC	4.16±0.45	14.76±0.95
HMC*	6.24±0.45	12.65±0.01
VI	3.94±0.01	12.77±6.11
VI*	3.84±0.01	12.32±6.36
LLA	8.85±2.09	23.84±3.54
Monte Carlo Dropout	7.88±2.81	25.75±4.44
Ensembles	7.69±0.27	24.41
MAP	5.76	

Table 4: Comparison of Bayesian Techniques Mohan and Scaife 2024 Page 6

However, the purpose of uncertainty quantification, as mentioned above, is not to improve accuracy but to provide a deeper insight into classifications and clarity to how the confidence of the model in its predictions Tyralis and Papacharalampous 2024. All the models discussed which used uncertainty quantification provided valuable additional insight into galaxy classifications, which can be used to inform decision making on these classifications, particularly on less certain and borderline cases Marusich et al. 2024.

5 Explainability in ML models

One of the main reasons ML models, particularly NN models, are considered untrustworthy is because there is a lack of understanding of how the models make their decisions Atemkeng et al. 2025. More basic models, called white box models, are inherently more interpretable because of how they make their decisions can be understood by viewing the model structure Cao et al. 2025. In more complex black box models, such as CNNs which are commonly used in computer vision systems like galaxy classification, it is difficult to interpret how decisions are made because of the complexity of the model Goh et al. 2025. As a result, to interpret these models separate explainability techniques must be used. The following section will outline modern ML explainability techniques in galaxy morphology classification tasks.

5.1 Explainability techniques used in classification systems

In the literature in this review, inherently white box methods are less common, as is the case in most modern galaxy image classification systems Cao et al. 2025. One of the white box models used Principal Component Analysis (PCA)

as a method of dimensionality reduction, finding a set of n vectors which describe the most variance in the original dataset. These vectors were then used to find similarities and outliers in different galaxy types Çakir and Buck 2024. Other models also used manual feature extraction techniques, which are also inherently explainable since the methods of extracting features are clearly defined Sarkar, Palit, and Bhattacharya 2024.

Most of the other techniques used are black box methods, using CNNs for automated feature extraction or other non-inherently interpretable methods. However, some of these papers used different techniques with the specific goal of interpreting the models. One model with features derived from pre labelled attributes used SHapley Additive exPlanations (SHAP) to interpret the importance of each of the features at a global and individual level Aguilar-Argüello et al. 2025. Another model used Local Interpretable Model-agnostic Explanations (LIME) to identify what areas of the images had the most importance, by changing pixel values in sections of the images at random and analysing how these changes affected predictions Goh et al. 2025. One paper also compared different saliency methods, to assess which parts of the image are relevant for class prediction Atemkeng et al. 2025. 10 saliency methods were used:

- Vanilla Gradient
- Smooth Grad
- Fast XRAI
- Vanilla Integrated Gradient (IG)
- Blur Integrated Gradient
- Smooth Blur Integrated Gradient
- Guided Integrated Gradient
- Grad-CAM
- Grad-CAM++
- Score-CAM

Atemkeng et al. 2025

5.2 Comparison of explainability techniques

Inherently interpretable explainability techniques are generally preferable than black box methods for a number of reasons, including:

- Less computationally demanding as they do not require additional computation for explainability
- More reliable as black box methods are approximations

- Less susceptible to effects of noise
- Interpretations are easier to act on, as they are defined by clear rules

Retzlaff et al. 2024

However, due to the greater complexity of black box models, they are often better at performing more complex tasks, such as image classification or feature extraction Goh et al. 2025. White box methods are inherently interpretable because of their simplicity, but simplicity also leads to performance drawbacks Cao et al. 2025.

In the literature in this review, extraction of manual features such as asymmetry metrics Sarkar, Palit, and Bhattacharya 2024 and other white box methods such as PCA Çakir and Buck 2024 are the most interpretable, as the methods are defined by clear rules. These methods also do not require any additional compute for interpretation. Among the other papers, SHAP Aguilar-Argüello et al. 2025 and LIME Goh et al. 2025 are among the most interpretable, however both are computationally intense Retzlaff et al. 2024 and LIME is only usable for individual analysis, being unable to provide global interpretability Goh et al. 2025. Other papers used less computationally intensive methods of explainability, like Smooth Grad, Fast XRAI, Blur IG Atemkeng et al. 2025. However, most of these methods were also less reliable and more sensitive to noise, with Grad-CAM and Score-CAM being the most reliable , although Score-CAM is also more computationally intensive Atemkeng et al. 2025.

6 Limitations in the current literature

Most of the literature in this review focused on model performance while often neglecting model explainability Serrano-Pérez, Díaz Hernández, and Sucar 2024. This is especially an issue considering there was also a general lack of manual feature extraction techniques, leading to entirely black box methodologies with unexplainable models using unknown features Huang, Wu, and Huang 2024. Some papers also used somewhat unreliable datasets, such as those labelled by members of the public which did not implement methods to correct biases associated with unprofessional labelling Sarkar, Palit, and Bhattacharya 2024. Additionally, there was a lack of literature which use hierarchical classification Semenov et al. 2025, particularly on a data set with labels which follow a hierarchy Conselice 2014. Considering the scientific applications of these classifiers, and the strong similarities between classes, there was also a distinct lack of uncertainty quantification techniques among the literature in this review Aguilar-Argüello et al. 2025.

7 Conclusion

This literature review discussed current galaxy morphology classification techniques, including feature extraction and dimensionality reduction methods, hi-

erarchical architectures for classification, and uncertainty quantification and explainability approaches used in galaxy image classification.

The most common feature extraction method identified in the reviewed literature was automatic extraction using deep learning models, specifically Convolutional Neural Networks (CNNs), as demonstrated by Serrano-Pérez, Díaz Hernández, and Sucar 2024, Huang, Wu, and Huang 2024, Mohan and Scaife 2024, Goh et al. 2025, Atemkeng et al. 2025, Semenov et al. 2025, Sarkar, Palit, and Bhattacharya 2024, Yin et al. 2024, and Feuer et al. 2024. However, it was observed that combining manual and automatic feature extraction generally yielded the best results, as noted by Sarkar, Palit, and Bhattacharya 2024 and Semenov et al. 2025. Image selection played a lesser role in performance improvements, likely due to the extensive preselection processes applied in datasets such as Nair and Abraham 2010 and Baillard et al. 2011, which primarily included the brightest images and discarded those without galaxies. Nonetheless, Huang, Wu, and Huang 2024 did remove underrepresented classes. More emphasis was placed on dimensionality reduction techniques such as greyscaling Semenov et al. 2025, removal of non-galaxy pixels Sarkar, Palit, and Bhattacharya 2024, and normalization of pixel values Dias 2025.

Morphological galaxy classifications inherently follow a hierarchical structure, with several possible hierarchies represented in models by Serrano-Pérez, Díaz Hernández, and Sucar 2024 and Aguilar-Argüello et al. 2025. Serrano-Pérez, Díaz Hernández, and Sucar 2024 implemented a Bayesian Convolutional Neural Network that accounted for major classes when predicting minor classes, while Aguilar-Argüello et al. 2025 employed several XGBoost models to represent class hierarchies. Both approaches performed as well as or better than baseline flat architecture models.

Uncertainty quantification was implemented in several papers, including Bayesian methods by Serrano-Pérez, Díaz Hernández, and Sucar 2024, Tyralis and Papacharalampous 2024, and Zhou et al. 2025, fuzzy C-means clustering by Marín Díaz, Gómez Medina, and Aijón Jiménez 2024, and a comparative study by Mohan and Scaife 2024. These studies found that models incorporating uncertainty quantification outperformed others, with variational inference and Hamiltonian Monte Carlo methods performing best according to Mohan and Scaife 2024. Additionally, all uncertainty methods provided a better understanding of model confidence, offering greater insight into classification decisions Marusich et al. 2024.

While some inherently explainable techniques such as manual feature extraction Sarkar, Palit, and Bhattacharya 2024 and principal component analysis Çakir and Buck 2024 were used, most explainability approaches relied on independent post hoc methods like SHAP Aguilar-Argüello et al. 2025 and LIME Goh et al. 2025. Though these methods provided in-depth insights into relevant image features, they were more computationally intensive than alternatives such as Grad-CAM and Smooth Gradients Atemkeng et al. 2025. Moreover, these less computationally demanding methods were found to be less reliable and more sensitive to noise compared to the more intensive techniques Atemkeng et al. 2025.

References

Aguilar-Argüello, G et al. (2025). "Morphological classification of galaxies through structural and star formation parameters using machine learning". In: *Monthly Notices of the Royal Astronomical Society* 537.2, pp. 876–896.

Ali, Ibrahim, Khaled Wassif, and Hanaa Bayomi (2024). "Dimensionality reduction for images of IoT using machine learning". In: *Scientific Reports* 14.1, p. 7205.

Atemkeng, MT et al. (2025). "A benchmark analysis of saliency-based explainable deep learning methods for the morphological classification of radio galaxies". In: *arXiv preprint arXiv:2502.17207*.

Baillard, Anthony et al. (2011). "The EFIGI catalogue of 4458 nearby galaxies with detailed morphology". In: *Astronomy & Astrophysics* 532, A74.

Çakir, U and T Buck (2024). "MEGS: Morphological Evaluation of Galactic Structure-Principal component analysis as a galaxy morphology model". In: *Astronomy & Astrophysics* 691, A320.

Cao, Yuzhou, Lei Feng, and Bo An (2024). "Consistent hierarchical classification with a generalized metric". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 4825–4833.

Cao, Zhuo et al. (2025). "Galaxy Morphology Classification with Counterfactual Explanation". In: *arXiv preprint arXiv:2510.14655*.

Conselice, Christopher J (2014). "The evolution of galaxy structure over cosmic time". In: *Annual Review of Astronomy and Astrophysics* 52.1, pp. 291–337.

Dias, Ian Alton (2025). "Combining Clustering and Classification methods for Galaxy Morphology Identification". PhD thesis. Dublin, National College of Ireland.

Feuer, Benjamin et al. (2024). "Select: A large-scale benchmark of data curation strategies for image classification". In: *Advances in Neural Information Processing Systems* 37, pp. 136620–136645.

Goh, Kam Meng et al. (2025). "An Interpretable Galaxy Morphology Classification Approach Using Modified SqueezeNet and Local Interpretable Model-agnostic Explanation". In: *Research in Astronomy and Astrophysics* 25.6, p. 065018.

Huang, Ruijie, Haoran Wu, and Jiayi Huang (2024). "Galaxy classification based on deep learning". In: *Proceedings of the International Conference on Image Processing, Machine Learning and Pattern Recognition*, pp. 577–582.

Marín Díaz, Gabriel, Raquel Gómez Medina, and José Alberto Aijón Jiménez (2024). "Integrating fuzzy c-means clustering and explainable ai for robust galaxy classification". In: *Mathematics* 12.18, p. 2797.

Marusich, Laura R et al. (2024). "Using AI uncertainty quantification to improve human decision-making". In: *arXiv preprint arXiv:2309.10852*.

Miranda, Fábio M, Niklas Köhnecke, and Bernhard Y Renard (2023). "Hiclass: a python library for local hierarchical classification compatible with scikit-learn". In: *Journal of Machine Learning Research* 24.29, pp. 1–17.

Mohan, Devina and Anna MM Scaife (2024). "Evaluating Bayesian deep learning for radio galaxy classification". In: *arXiv preprint arXiv:2405.18351*.

Nair, Preethi B and Roberto G Abraham (2010). "A catalog of detailed visual morphological classifications for 14,034 galaxies in the sloan digital sky survey". In: *The Astrophysical Journal Supplement Series* 186.2, p. 427.

Retzlaff, Carl O et al. (2024). "Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists". In: *Cognitive Systems Research* 86, p. 101243.

Sarkar, Ankita, Sarbani Palit, and Ujjwal Bhattacharya (2024). "Demystifying galaxy classification: An elegant and powerful hybrid approach". In: *2024 39th International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE, pp. 1–6.

Semenov, Vasyl et al. (2025). "Galaxy morphological classification with manifold learning". In: *Astronomy and Computing* 52, p. 100963.

Serrano-Pérez, Jonathan, Raquel Díaz Hernández, and L Enrique Sucar (2024). "Bayesian and convolutional networks for hierarchical morphological classification of galaxies". In: *Experimental Astronomy* 58.2, p. 5.

Szeliski, Richard (2022). *Computer vision: algorithms and applications*. Springer Nature.

Tegmark, Max et al. (2004). "Cosmological parameters from SDSS and WMAP". In: *Physical review D* 69.10, p. 103501.

Theng, Dipti and Kishor K Bhoyar (2024). "Feature selection techniques for machine learning: a survey of more than two decades of research". In: *Knowledge and Information Systems* 66.3, pp. 1575–1637.

Tyralis, Hristos and Georgia Papacharalampous (2024). "A review of predictive uncertainty estimation with machine learning". In: *Artificial Intelligence Review* 57.4, p. 94.

Vázquez-Mata, JA et al. (2022). "SDSS IV MaNGA: visual morphological and statistical characterization of the DR15 sample". In: *Monthly Notices of the Royal Astronomical Society* 512.2, pp. 2222–2244.

Yin, Lirong et al. (2024). "Convolution-Transformer for Image Feature Extraction." In: *CMES-Computer Modeling in Engineering & Sciences* 141.1.

Zhou, Xingchen et al. (2025). "Estimating photometric redshifts for galaxies from the DESI Legacy Imaging Surveys with Bayesian neural networks trained by DESI EDR". In: *Monthly Notices of the Royal Astronomical Society* 536.3, pp. 2260–2276.