Galaxy Morphology Classification: A Survey

Aditi Ravishankar
Department of Electronics and and Communication Engineering
RNS Institute of Technology
Bangalore, India
aditiravishankar21@gmail.com

Aishwarya K Karanth
Department of Electronics and Communication Engineering
RNS Institute of Technology
Bangalore, India
aishwaryakaranth27@gmail.com

Adhishreya P
Department of Electronics and Communication Engineering
RNS Institute of Technology
Bangalore, India
adhishreyas@gmail.com

Ameena
Department of Electronics and Communication Engineering
RNS Institute of Technology
Bangalore, India
ameenaameena464@gmail.com

Apoorva N Hegde
Department of Electronics and Communication Engineering
RNS Institute of Technology
Bangalore, India
apoorvanh@gmail.com

***Abstract*—To understand the large-scale structure of the universe, studying the morphology of galaxies is imperative. Scientists build a deeper understanding of the formation and evolution of galaxies by classifying them into different categories based on morphological parameters. With evolving galaxies and growing galaxy image data it was impossible for manual sorting. Automated means were then considered and newer algorithms evolved for feature extraction, reduction and classification. In this paper, we present a survey of 20 research efforts that employ various machine learning and deep learning techniques for the classification of galaxies based on morphology to gather insights into the accuracy that is obtained by these approaches. The pre-processing methods used on the data, models employed and the performance achieved in each work under study are surveyed in this paper.**

***Index Terms*—Galaxy Classification, Galaxy Morphology, Deep learning, Machine learning, Image Processing**

I. Introduction

13.8 billion years ago, the Big Bang led to the formation of the universe that we know today, comprising of billions of galaxies of sundry sizes and shapes. Galaxies are gravitationally bound collections of interstellar dust, gas and stars and these mysterious creations are a fundamental topic in cosmology. Galaxies exhibit a plethora of different morphologies and knowledge of the same aids astrophysicists to test existing theories pertaining to the nature of the universe and construct conjectures of their own. Previously, astronomers would use manual techniques for the purpose of classification of galaxies but with the advent of epochal fields like Machine learning and Deep learning, this process has since been automated, which not only greatly reduces the time consumed but also prevents misclassifications and inaccuracies that are inevitable in manual methods. There has been an explosion in the astronomical data available in recent times and the manual analysis of the same is not feasible. Morphological classification of galaxies is the process of categorising galaxies based on visual shapes. The work surveyed in this paper exhibits different styles of classification. In [1], [6] and [8], there was a three-category classification - elliptical, spiral and irregular. In [2], they attempted a five-category classification - spiral, elliptical, round, disk and other. In [3], they tried a three-category classification - elliptical, spiral and irregular initially and then did a 7-category classification based on the Hubble scheme. Another attempt [4] made use of the Galaxy Decision Tree (Willett et al., 2013) in which there are 11 questions and they have to predict the probabilities of 37 answers. In [5], the authors discuss the importance of Shape descriptors that could be useful indicators in classifying galaxies. The motivation for presenting this survey stems from the need to find the most optimal, efficient and accurate existing method for the classification of galaxies based on morphology. In [7], two approaches for galaxy morphology classification were proposed - one based on non-parametric morphology and traditional Machine learning algorithms and another based on Deep learning, and these were used to perform a classification with three categories - spiral, elliptical, barred spiral. In [9], the categories of classification were 5 classes - completely round, smooth, in-between smooth, cigar-shaped smooth, edge-on and spiral. In [10], models were used to differentiate between spiral galaxies, elliptical galaxies and star/unknown galactic objects. The Sloan Digital Sky Survey is a sky survey which consists of deep, multi-color images covering more than a quarter of the sky and has created 3-dimensional maps containing more than 930,000 galaxies and more than 120,000 quasars [13]. The Galaxy Zoo 2 project involved a crowdsourcing effort via a web based platform to classify the galaxies as per a catalogue containing 11 questions and 37 different possible responses in all. In [11], they eliminated certain questions from the catalogue such as number and tightness of spiral arms. Deep Learning is said to extract the most relevant features automatically using non-linear transformations. In [12], self-supervised learning technique works on unlabelled data with certain attributes that can be derived from the images. In [16],

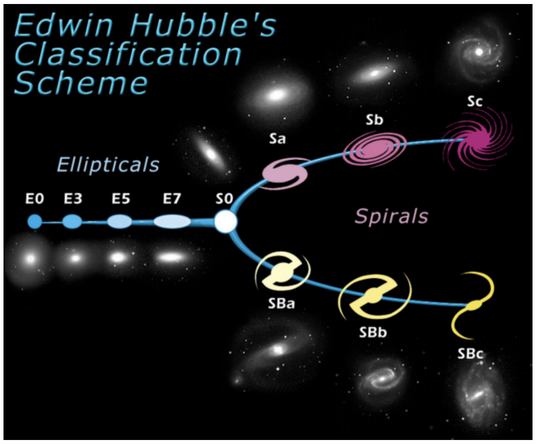

Fig. 1. [2] [5] Hubble galaxy classification scheme

classification is done by considering 13 parameters measured by machine (ESO-LV) into five types (E, S0, Sa+Sb, Sc+Sd and Irr). In [19], they have classified the images of radio galaxies considering morphological parameters. They have taken into account Fanaroff-Riley (FR) class of radio galaxies and bent-tailed morphology. In [20], Hubble's classification scheme is used to classify the galaxies.

Our motivation for this paper was to find approaches to the galaxy classification problem with good performances to aid further research. We collected papers, filtered out 20 of them and analysed each paper based on the following questions -

1) What were the sources of the data being used?
2) What were the classes/ labels being considered?
3) What were the preprocessing and data augmentation techniques being employed to enhance the image data?
4) What was the method (ML/DL models) being used for the classification process?
5) What was the overall performance of the model?
6) What could be the reasons for the performance not being better and what are the suggestions for improvement?

II. Design Methodologies

A. Datasets

In this subsection, we introduce the sources of data made use of in all the related works surveyed. There are numerous large datasets available containing thousands of images of galaxies. Most of the datasets contain galaxies imaged by the Sloan Digital Sky Survey (SDSS).

In [1], an attempt was made to manually collected their images from Google using a browser extension which led them to do a significant amount of pre-processing, discussed in the following subsection. Some of this pre-processing could be avoided by using datasets like the one by Galaxy Zoo. The current Galaxy Zoo combines images from the SDSS with the most distant images yet from Hubble's CANDELS project. The GZ data has fuelled multiple papers. The GZ data comes with 61578 pre-classified 424 × 424 RGB images of galaxies taken from SDSS. It also comes with the classifications that have been gathered from a crowdsourced quiz, given in the form of probabilities of the answers to 37 questions. This dataset was used in [2], [4] and [9].

In [3], a dataset obtained from Zsolt Frei's galaxy catalogue was used. Hubble Tuning Fork images were used as model or prototype images in [5]. In [6],Sloan Digital Sky Survey (SDSS) Data set was used. In [7],the main dataset was employed from Sloan Digital Sky Survey Data Release 7 (SDSS-DR7) and Galaxy Zoo catalogues with 670,560 galaxies. In [8],the dataset used was taken from the EFIGI catalogue This dataset consists of more than 11,000 images. In [9],the galaxy images were taken from the galaxy zoo dataset which contains 61578 JPG colour galaxy images where each image has a size of 424x424x3 pixels. In [10],full sample sets of Galaxy images from SDSS DR7 was used. The dataset has 667,935 entries and each correlates with an object in the SDSS database. After some sort of filtering, the dataset was reduced to 251,867. [11] uses Galaxy Zoo 2 and Nair et al. catalogues 2010 for training .For testing the catalogues used were Huertas-Company et al. 2011 and Cheng et al. 2011 catalogues. Southern Photometric Local Universe Survey (S-PLUS) is an imaging survey which uses 12 optical bands. [12] uses the first release of the S-PLUS which includes both images and catalogues of detected objects and the labels for this dataset are obtained by matching the astronomical coordinates from S-PLUS to SDSS spectra catalogue [13] uses Galaxy Zoo 2 catalog and datasets derived from Catalog Archive Server of SDSS(Sloan Digital Sky Survey) where the image size was 120x120 px. An approximate of 2,45,609 images were used. Zsolt frei Catalog comprises of nearly 113 images from nearby galaxies taken in multiple pass bands the images had high resolution and had better calibrations therefore extensive pre-processing was not required [14].

B. Preprocessing

This Subsection introduces different processing techniques used to enhance desired features present in an image which are relevant for further analysis.Image processing is very essential to analyze and make important inferences about celestial objects. It plays a vital role in analyzing and interpreting galaxy images and to improve and analyze the properties of celestial objects. Some image processing steps were involved before feeding the image data into the models. This is done to improve the image data and suppress the undesired distortions.

- Image processing played a vital role in [1] as their work was focused on comparing the performance of the classifier with and without using image processing techniques. They had to resize their images as the data was collected from Google with varying sizes. The different image processing methods using which they tested their classifier were Canny Edge detection, Histogram Equalization, and Median Filtering.In [2], they began with cropping their images to remove secondary objects and rotating them to make the principal axes vertical, following which they performed background subtraction Since there were a huge number of features, to make their model less computationally intensive and reduce overfitting, Principal Component Analysis was performed to reduce the number of features while also not allowing the loss of a lot of galactic information.

- In the image processing phase of [3], they scaled, rotated, cropped and centered the images. In the Feature Extraction phase they measured 6 morphic features for each image considering visual characteristics and also generated PCA features for each image. [4], they focused on studying the importance of scale in the process of Galaxy Image Classification. They performed random rotation, horizontal and vertical flipping and shifting for the purpose of Data Augmentation. They pre-processed the images to have three different scales - a normalized scale such that the whole galaxy is shown in the image, 6464 and 256256 and proceeded with determining which scale would produce better results.
- [5] discusses the importance of Shape descriptors that could be useful indicators in classifying galaxy shapes. The performed thresholding using Otsu's method, morphological opening operation and flood-fill operation to fill the holes on the images in the preprocessing stage. In [6],they focused on preprocessing techniques like histogram equalization,contrast stretching and noise filtering which is very helpful in feature extraction, image analysis and image display. In [6] and [8], some techniques like rotation, reflection and cropping were applied to the training data for augmenting the image database. As a result of image augmentation, overfitting was avoided.
- In [7], Multilayer Perceptron models were used to classify the galaxies. Multiple layers of non-linear transformations were used. For preprocessing in [9], they cropped the image in the range of [170,240] at the first step, which allowed all the main information to be contained in the center of image, and also eliminates many noises like other secondary objects. Then, the image was resized to 80x80x3 pixels, which is just dimension reduction and randomly rotated with 0, 90, 180, 270 degrees and randomly horizontally flipped. Adjustments were made to the brightness, contrast, saturation and hue, followed by image whitening. In [10], independent component analysis was performed to determine the most significant components which is necessary for classification.
- Image Resizing: The images used might be of different sizes, while building models it is often required that all the images have a fixed size, therefore resizing images are carried out before training [14]. The images are down sampled to 49x49x3 as it reduces the computation time on each image [11]. Initially, coloured image is converted to a grey scale image where the pixel intensities vary from 0-255. A binary image has only two pixel values. [15] Otsu transform is one of the image binarizing techniques which involves an iterative process which separates the foreground and background pixels [13] [15].
Data Augmentation: It is a form of artificially expanding the data-set. It trains the model on a variety of images thereby allowing it to generalize well. Certain manipulations on the images are considered including rotation [14] [11],flipping the images horizontally and vertically [11],zooming the images in and out [11],shifting -horizontal or vertical shifts and centering [14].
- In [17], they first generated images that are independent of orientation and scale and then used principal component analysis for a more compact representation of images. [18] presents various methods in which astronomy images can be processed. According to [18], linear filters remove noise efficiently but there is a trade-off that is the texture of the image is not preserved which is why non-filters are used. The non-linear filter used in [18] is non-local means filter. Non-local means filter replaces the intensity value of the target pixel with an average of a selection of intensities of other pixels where small regions centered on other pixel is compared to the region having the target pixel as its center, averaging here is performed when both the regions have a high rate of similarity and therefore the texture and details are preserved. Object segmentation is crucial in image processing as astronomical telescopes often capture overlapping celestial objects. Chan Vese algorithm is used for segmentation [18]. Segmentation algorithm is used for segmenting objects lacking distinct boundaries. Watershed algorithm is a segmentation algorithm used to plot the representation of distance mapping between galaxies in [18]. To prevent diffusions across high gradients, Random Walker algorithm is used [18]. Watershed algorithm is a segmentation algorithm used to plot the representation of distance mapping between galaxies in [18]. For preprocessing in [20], alpha-rooting transform, heap transform and Paired (Grigoryan) Transform are used. For galaxy detection, background subtraction and nonlinear filtering was applied for the segmentation of galaxies, followed by edge detection. In [16], they found that filtering the images of galaxies in the training set improved the classification.

C. Models

The goal of this sub-section is to provide insight into the techniques, architectures and models that we encountered in the papers surveyed in this work. Many algorithms of Machine learning and Deep learning have been used. Machine learning is a field of research that deals with learning algorithms used to parse and learn from data and make predictions based on the same. Deep Learning comes under Machine learning and is concerned with algorithms that create artificial neural networks to make decisions. Convolutional Neural Networks are types of deep neural networks which are extensively used for image recognition and classification.

- Feature extraction: WND-CHRM is an image analysis technique which works by extracting large volumes of image feature descriptors. [13] used the technique and different feature descriptors are required for identifying unique features and around 2883 features were reduced

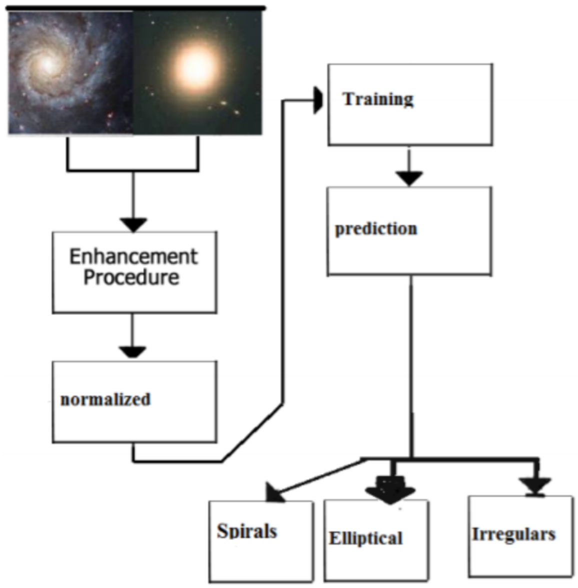

Fig. 2. Non negative matrix classification scheme [14]

by means of a Fischer score that was set to a certain acceptable score. Classification was performed using the Weighted Nearest Neighbour scheme considering the Fischer scores as weights. For each question from the catalogue, samples were divided into training and test set making sure that the number of samples for each class remains equal.

- Non-negative matrix factorization is a dimensionality reduction technique where the given matrix can be expressed as the product of two other matrices, the basis matrix comprising of the image features and the coefficient matrix. In [14] both the test and training samples were normalized While training the basis matrix(W) and the coefficient matrix (B) were adjusted such that their product gives the original database vector (V) considering the Euclidean norm to reduce the distance. For testing the test set S was represented by product of coefficient A and W matrices. The label will be the maximum coefficient value of A. [15] extracted different features from binary images such as detecting the circularity , finding the view angles, determining the existence of bulge, detecting the tightness and number of spiral arms. A knowledge base is constructed with CBH (Case Based Reasoning) constructed from human-annotated data .CBCVH builds on extracted features from images comprised of numerical values for each feature. RBCVF rule based system constructed from the extracted features, which comprised of two parts first the feature of the subclass, second the next rule that must be fired. In [15], for few features weights were assigned to each feature considering how close the new cases should be to the training cases. For the remaining features minimum value is chosen from the new case and cases in the knowledge base and then class was assigned to them as per case base . [12] Self supervised learning involves extracting attributes easily extractable from the data in pretext task and use those for downstream tasks.
- In [5] after the Elliptic Fourier Descriptors were obtained and Principal Component Analysis was performed on the EFDs, the Euclidean distance between a particular image under consideration and all the model images was calculated. The average distance was computed and the minimum value was considered as the best match.
- In [17], the classification of galaxies based on morphology is done using three different machine learning algorithms on three, five, and seven galaxy types. The Hubble tuning fork scheme was used to classify the galaxies as spiral, elliptical and irregular. Naive Bayes, rule-induction algorithm C4.5, and random forest were used. In the ensemble method used here where an ensemble consists of a number of classifiers whose individual decisions are made in a certain way usually by voting to classify new examples. The ensemble method used here is bagging.
- In [2], they performed classification using SVM with an RBF kernel, decision tree, random forest, AdaBoost classifier, and KNN. They decided to proceed with regression by predicting the probability density function of the classification, to account for uncertain classifications. They also applied three clustering algorithms to the PCA dataset - k-means clustering, Agglomerative clustering with complete linkage and ward linkage.
- In [3], they trained SVM with RBF kernel, Random Forest, Naive Bayes classifiers and their AdaBoosted versions on the training subsets. They implemented 10-fold cross validation (3 times) on the training set, finding the PC vector basis each time and they gave the average of these 3 runs as their overall accuracy and standard deviation for each algorithm. For each algorithm, they classified the galaxies into 3 classes and 7 classes, first using only morphic features, then using only PCA features, and finally using both morphic and PCA features.
- In [10], Reliable machine learning models were used for distinguish between spiral galaxies, elliptical galaxies or star/unknown galactic objects.Different algorithms like Decision tree lea and fuzzy inferencing systems were applied and Random Forest classifier with number of trees as a particular sample While processing a specific sample, the output by each of the individual trees was considered and the resulting node was taken as the final classification.
- In [20], the classification of galaxies was considered in two parts, the first part being ellipticals and spirals and the second being simple spirals and barred spirals. Sparse representation of data was considered where sparse representation decomposition is achieved by optimizing an objective function that includes both a reconstructive and a discriminative term. The goal of the data decomposition is to achieve the sparsest representation, that is some coefficients have a large value, and most of them have

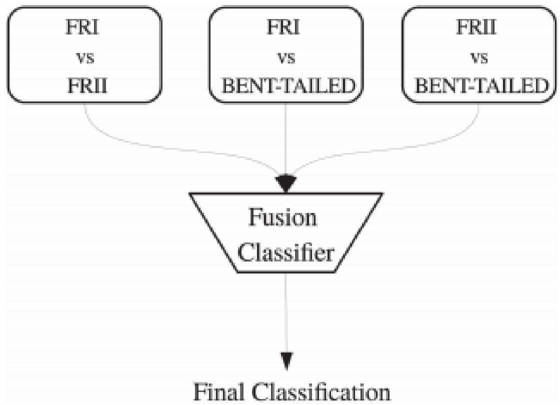

Fig. 3. [19] Fusion model with majority voting ensemble classifier, which combines the predictions from the three binary classifier models to make the final prediction. The figure shows the individual predictions from the three binary classifiers being fed into a fusion classifier, which gives the final classification. This model is ideal in situations where the individual models have a slight bias and is also beneficial to identify odd inputs.

values nearing zero. For the first part, 6 features were used for training, and for the second part, only the asymmetry index was trained.

- In [16], the classification of galaxies is done by an artificial neural network algorithm. The neural network is here behaves as a Bayesian classifier. Backpropagation algorithm is used to train a network on part of the catalog, and then using the measured parameters, predict the classification for the rest of the catalog. They considered five types(E, S0, Sb, Sc+Sd and Irr) of galaxies with 13 parameters. Bayesian classifier here assigns a posterior probability to five classes considered. The threshold activation function used here was the sigmoid function. Weights were determined by minimizing least-squares which was done in backpropagation using gradient descent. The learning rate and momentum were constant. A total of 1.5M iterations were done.
- In [19], the radio galaxies are classified based on morphology using convolutional neural networks(CNN). It considers Fanaroff-Riley(FR) class of radio galaxy and bent-tailed morphology. CNNs classified images of the FRI and FRII and bent-tailed radio galaxies.
- In [8], Deep learning approaches were used for performing classification. The Hubble galaxy classification scheme was used to classify the galaxies as spiral, elliptical and irregular. The CNN uses two layers such as convolutional layer and pooling layer. The input to a convolutional layer was an image, the outputs of which are resulted in feature maps that are convolved with weighted filters and ReLu activation Funcion that is non linear in nature which is then applied to the weighted sum of these convolutions in order to produce a final feature map. The optimization technique used was Gradient Descent which updated all the weights at once after iterating through all the samples in the training dataset. Back Propagation Algorithm was then used to train the network then classification output was predicted.
- In [1], a deep convolutional neural network called Inception module was utilised with a global average pooling replacing the fully connected layer to reduce the number of parameters thereby reducing the computation requirements. They passed the images processed by the three methods mentioned in the previous section, along with their unprocessed equivalents to compare the classifier's performance.
- In [4], CNNs were used with 9 convolutional layers and 3 pooling layers in the convolution segment, followed by dropout (with a dropout rate of 50 %) and maxout layers without which they claim the training would be a problem due to overfitting. A single neural network was used for training and predicting the probability of each of the 37 answers of the decision tree.
- In [5] after the Elliptic Fourier Descriptors were obtained and Principal Component Analysis was performed on the EFDs, the Euclidean distance between a particular candidate image and all the model images was calculated. The average distance was computed and the minimum value was considered as the best match. [11] used CNN with Keras library. It had 4 convolutional layers, drop out to avoid overfitting and max pool layers. For GZ2 catalog, the model was trained on a binary classification model for each question (disk/flat, edge-on, bulge prominence, roundness, merger). An agreement parameter was devised to specify the consistency for classification. A T-Type model trained on N10 catalog assigned an integer to each galaxy,(-3 to 10) and used mean squared error (mse)for loss.The transitions of types from E to spiral morphology is observed.E11 and S0 are further separated using pretrained weights from the previous process, barred galaxies are further separated as strong,intermediate weak N10 bars using weights from GZ2 model.
- In [12], Source Extractor was used to obtain magnitude values from images. They scaled them by dividing the values by 30 and used them as targets for prediction as a part of the pretext task. In [12] VGG-16 architecture yielded a better result both with pretrained on ImageNet weights and training from scratch .2048 unit fully connected layer with dropout of 0.5 was used. For the classification task, the output was passed through a softmax layer and the loss function was set to categorical cross-entropy and for the regression task-modified relu and Mean absolute error were used. Feature extraction and fine-tuning were done freezing some layers and training the remaining.
- In [9], a type of residual networks (ResNets) for galaxy morphology classification was proposed where, by visualizing filters weights and feature maps the first layer filters detect the different galaxy edges, corners, etc. from the original pixel, then the edges are used to detect shapes in second layer filters, such as bar and elliptical, and then these shapes are used to detect more advanced features

in high level layer filters. Different filters learn different types of color information, mainly red and blue, which might correspond to the color of the galaxies.

- In [6], Morphological Galaxy Classification was done using Machine Learning Algorithms And image analysis method where feed forward neural network and locally weighted regression methods were used for performing classification task to classify the galaxy into three categories namely elliptical, spiral, irregular.The raw data images were used as input without any expert knowledge in feature designing or optimization of segmentation parameter.For Feature Extraction architecture convolutional layer was proposed where feature detectors were used which basically serves as neural network filters and an two principles fully connected layers for the classification. Relu activation function(non linear activation function) was used as it simplify computation.
- In [7], two approaches were used for achieving classification, one was non-parametric morphology and traditional machine learning algorithms and the second was a deep learning approach. Decision Tree of various versions like CART(classification and regression tree) and support vector machines were used. Non linear transformation were performed in different layers to build a Deep neural network. They also performed two class (elliptical and spiral) and three class classification (elliptical,spiral,irregular).

III. Performance of various methods

This section is about the performance of the different models utilised in the works surveyed. Regarding the methods used to evaluate the performance, various metrics have been employed by the authors.

In [1], the best results were seen for unprocessed images. The processing techniques could have meddled with the image quality and led to the loss of some important features. Out of the three processing techniques, histogram equalization showed the best results and Canny edge performed the worst. The results were taken for different numbers of epochs with the average accuracies for all the methods being 74.2% for unprocessed images, 57.2% using histogram equalization,56.3% using median filtering and 53.1% using Canny edge detection. The highest accuracy obtained was 78.3% on unprocessed images at 80-90 epochs. The authors of [2] found the best performance in the random forest algorithm with a 67% classification accuracy. Their regression approach was more successful with a 95% accuracy using the random forest regressor. In [3], random forest with only morphic features performed the best. The ideal number of PCA features was found to be between 8 and 24 and classification using only morphic features performed better. It was concluded in [4] that normalizing the scale of the galaxy image produces better results with regard to the mean square error value, followed by 256256 scale and then 6464 scale. In [5], 42 images out of 50 matched with Hubble's scheme with the overall agreement being 84%.

In [6] ,Deep Learning has achieved significant results and a huge improvement in visual detection and recognition. The best performing models were found to be Random forest and Neural networks. In [7],the results obtained considering two classes are very consistent ($OA \geq 98.7\%$) and good for the three classes problem (Overall accuracy 82%). Both Deep and Traditional Machine Learning approaches have over 94.5% overall accuracy while classifying galaxies into two classes (elliptical and spiral). Considering the two class-separation case, 99% overall accuracy was achieved in average by using the deep learning models, and 82% when using three classes. In [8],the architecture was trained over 4238 images and achieved a 97.772% testing accuracy. In [9],the model achieves a high classification performance, the overall accuracy on testing set is 95.2083% and the accuracy of the 5 galaxy types are: completely round, 96.6785%; in-between, 94.4238%; cigar-shaped, 58.6207%; edge-on, 94.3590% and spiral, 97.6953% respectively. The average precision, recall, F1 and AUC of the model were 0.9512, 0.9521, 0.9515 and 0.9823. In [10], random forest gave the best results.

In [13], they considered human classification as the ground truth and achieved 85% performance for 8 features . [12] Comparisions were made between classification schemes for different datasets. For the StarGalaxy dataset, from scratch - 98%, extracting features from a model trained on ImageNet - 93%; extracting features from a model trained on the pretext task - 97%; fine-tuning a model trained on ImageNet - 99%; fine-tuning a model trained on the pretext task - 99%. [15] achieved a performance similar to that of human classification. In [14], a 93% of correct classification rate was achieved. [11]For GZ2 based model different level of accuracy is reached for different questions, 98% for Q1,97% for edge on,bar and merger.As of N10 based 96% for strong bar,80% for weak bars.

[17] used three machine learning algorithms- Naive Bayes, rule-induction C4.5 and random forest. Ensemble method called bagging was used here. Random forest yielded more accurate results compared to the other two algorithms that is 91.64% accuracy for three-class case, 54.72% for five-class case, and 48.62% accuracy for the seven-class case. Ensemble method achieved better results than individual classifiers. [20] SVM classifier was trained using 14 galaxies and validated with 3. The classification for part (1)ellipticals and spirals and (2)simple spirals and barred spirals was 100%. [16] used artificial neural network algorithm to classify the galaxies. The neural network behaved as Bayesian classifier. The accuracy was 64 percent when the highest probability was considered and the accuracy was 90 percent when the first or the second choice was considered. [18] used CNN for classification. The accuracy of bent-tailed radio galaxies was 95% and 91% for FRI and 75% for FRII classes.

IV. Conclusions

In this paper, we have performed a survey of research efforts on galaxy morphology classification. We have considered 20 papers and examined the datasets used, preprocessing techniques, design methodologies and the results obtained. In [1], unprocessed images had the best results, and amongst the processed ones, the ones which underwent histogram equalization had the best results because it works best on grayscale images. Better results could be obtained by the usage of other architectures of CNNs like VGG and expanding and refining the dataset (as the images were randomly collected from Google). In [2], the regression approach performed better and the unsupervised analysis that they performed showed that eccentricity and brightness constitute the major variation between the galaxies. In [3], the morphic features, being less prone to morphological variations were more effective than the PCA features. The features were not linearly separable, and the RF classifier seemed to be the most robust model. The authors claim that the accuracy of the morphic features was tampered by the cloud of dust and the background noise in the images and suggest the classification to be done in stages, first into 3 and then 7 categories for better results. The authors of [4] stressed on the impact of scale on the classification process and suggest incorporating scale normalization in larger neural networks for classification. In [5], the authors obtained an overall accuracy of 87% (AUC) and suggest that this can be improved by adding more training data. With the use of larger datasets, there is an increase in the processing power, memory size and GPU requirements and it was possible to train a deeper, larger and more complex models in [6]. Both deep learning and machine learning give high accuracies for two-class classification but this accuracy reduces when the number of classes are increased which can be dealt with by using K-fold cross validation technique such as in [7]. 5 trials were performed in [8] to improve the accuracy of the architecture for classifying galaxies using deep convolutional neural networks . In [9], the proposed model could be trained on bigger and high quality datasets to get better results. In [10], different algorithms were used to perform classification and a random forest with more number of trees could be used to yield better results. In [11] The authors had eliminated certain samples having poor acceptance scores and this resulted in the elimination of those data points with intermediate probability values. As in [12] self-supervised learning could be used when the labelled data is sparse or while performing unsupervised tasks like clustering. Papers that have used the Deep learning approach added dropout layers to reduce overfitting. The technique in [12] could not be applied for all data-sets like those extracted from SDSS and extracting properties other than magnitudes might make it work with different datasets. In [13], the accuracy of classifications changes as the threshold value changes. The authors of [14] suggest using data with less noise and overlapping points. In [15], feature extraction was an elaborate process and instead, a Deep learning algorithm like CNN which has automatic feature selection could be used. In [16], an ANN algorithm was used where the network behaved as a Bayesian classifier whose accuracy was 64%. The accuracy may be improved by introducing additional hidden layers and using different classifiers like random forest. In [17], although random forest gave highly accurate results for the three-class case, it gave less accurate results for five-class and seven-class case. Ensemble method yielded more accurate results than individual classifiers. In [19], CNN was used to classify radio galaxies and it yielded a highly accurate result. The accuracy could be improved by using different architectures. In [20], SVM classifiers yielded a high accurate result but only 3 images were used to validate. A more accurate result could be gotten by increasing the number of images for validation. Overall, Deep learning methods proved to be more successful than Machine learning models.

Our aim is that this survey would help researchers experimenting on galaxy morphology classification to gain insight into the repercussions of using the different datasets discussed, the preprocessing techniques and models in order to come up with a well-performing classifier that encapsulates the best features of all the related works.

References

[1] Wahyono, Muhammad Arif Rahman, and Azhari SN, Classification of galaxy morphological image based on convolutional neural network, Vol.5 , Issue 6 , June 2018.

[2] Alexandre Gauthier, Archa Jain, and Emil Noordeh,Galaxy Morphology Classification,2016.

[3] Siddhartha Kasivajhula, Naren Raghavan, and Hemal Shah,Morphological Galaxy Classification Using Machine Learning

[4] AbdulWahab Kabani and Mahmoud R. El-Sakka,How Important is Scale in Galaxy Image Classification?,

[5] Ishita, Dutta , S. Banerjee & M. De,Shape Descriptors in Morphological Galaxy Classification

[6] Gaurav Kiran Tiwari, Pooja N Mishal, Prof. Tejaswini Bhoye,Deep convolutional neural networks for galaxy morphology classification,Vol.6 , Issue 3 , March 2019

[7] P. H. Barchia, R. R. de Carvalho, R. R. Rosa, R. A. Sautter, M. Soares-Santos, B. A. D. Marques, E. Clua, T. S. Goncalves, C. de Sa-Freitas, T. C. Moura,Machine and deep learning applied to galaxy morphology

[8] Nour Eldeen M. Khalifa Mohamed Hamed N. Taha ,Aboul Ella Hassanien, I. M. Selim,Deep galaxy V2: Robust deep convolutional neural networks for galaxy morphology classifications

[9] Jia-Ming Dai,Jizhou Tong,Galaxy morphology classification with deep convolutional neural network

[10] Adam Gauc, Kristian Zarb Adam, John Abela,Machine learning for galaxy morphology classification

[11] H. Domnguez Sanchez,M. Huertas-Company, M. Bernardi, D. Tuccillo and J. L. Fischer,Improving galaxy morphologies for SDSS with deep learning,2018

[12] Ana Martinazzo, Mateus Espadoto, Nina S. T. Hirata,Self-supervised learning for astronomical image classification

[13]]. Joe George,Evan Kuminski, John Wallin, Lior ShamirCombining human and machine learning for morphological analysis of galaxy images

[14] I.M.Selim ,Arabi E. Keshk ,Bassant M.El ShourbugyGalaxy image classification using Non-Negative matrix factorization

[15] Devendra Singh Dhami, David B. Leake, Sriraam NatarajanKnowledge-Based morphologicalcClassification of galaxies from vision features

[16] Storrie-Lombardi, M. C., Lahav, O., Sodre, L., Jr., & Storrie-Lombardi, L. J.,Morphological Classification of galaxies by Artificial Neural Networks

[17] Jorge de la Calleja, Olac Fuentes, Automated Classification of Galaxy Images

[18] Diganta Misra , Sparsha Mishra and Bhargav Appasani,Advanced Image Processing for Astronomical Images

[19] A. K. Aniyan, K. Thorat, Classifying Radio Galaxies with the Convolutional Neural Network

[20] John Jenkinson, Artyom Grigoryan, Mehdi Hajinorooz, Raquel Díaz Hernández, Hayde Peregrina Barreto, Ariel Ortiz Esquivel, Leopoldo Altamirano, Vahram Chavushyan, Machine Learning and Image Processing in Astronomy with Sparse Data Sets