RAA 2021 Vol. **21** No. **10**, 264(8pp) doi: 10.1088/1674-14527/21/10/264

http://www.raa-journal.org *http://iopscience.iop.org/raa*

***R**esearch in*
***A**stronomy and*
***A**strophysics*

Automated classification technique for edge-on galaxies based on mathematical treatment of brightness data

Mohamed Eassa[1], Ibrahim Mohamed Selim[1,2], Walid Dabour[3,4] and Passent Elkafrawy[3,5]

[1] Computer Science Department, Integrated Thebes Institute, Cairo, Egypt; *eassa_h@yahoo.com*
[2] Faculty of Computer & Artificial Intelligence, Sadat City University, Sadat City, Egypt
[3] Math and Computer Science Dept., Faculty of Science, Menoufia University, Egypt
[4] Computer Science Dept., Taibah University, Alula Branch, Kingdom of Saudi Arabia
[5] School of Information Technology and Computer Science (ITCS), Nile University, Egypt

Received 2021 March 29; accepted 2021 July 26

Abstract Classification of edge-on galaxies is important to astronomical studies due to our Milky Way galaxy being an edge-on galaxy. Edge-on galaxies pose a problem to classification due to their less overall brightness levels and smaller numbers of pixels. In the current work, a novel technique for the classification of edge-on galaxies has been developed. This technique is based on the mathematical treatment of galaxy brightness data from their images. A special treatment for galaxies' brightness data is developed to enhance faint galaxies and eliminate adverse effects of high brightness backgrounds as well as adverse effects of background bright stars. A novel slimness weighting factor is developed to classify edge-on galaxies based on their slimness. The technique has the capacity to be optimized for different catalogs with different brightness levels. In the current work, the developed technique is optimized for the EFIGI catalog and is trained using a set of 1800 galaxies from this catalog. Upon classification of the full set of 4458 galaxies from the EFIGI catalog, an accuracy of 97.5% has been achieved, with an average processing time of about 0.26 seconds per galaxy on an average laptop.

Key words: techniques: image processing — methods: data analysis — galaxies: formation

1 INTRODUCTION

Astronomical studies of galaxy classification are gaining momentum to classify the detected huge number of galaxies, estimated to be in the range of 200 billion (Gott et al. 2005). These studies aim at investigating galaxy types and properties to shed light on the origin and developmental stages of our universe in general and our edge-on Milky Way Galaxy in particular (Sparke & Gallagher 2007). Exploring galaxy formation and evolution are central to modern cosmology (Khalifa et al. 2018). Figure 1(a) displays an edge-on galaxy (NGC 891) and Figure 1(b) displays a face-on galaxy (Messier 74). The visual appearance of galaxies, their morphologies, shapes and forms provide insight into their composition and their evolutionary history. Also, galaxies' chemical compositions were found to shed light on the formation and evolution of our universe. To facilitate these studies, morphological classification of galaxies was introduced (Elsayed Abd Elaziz et al. 2019).

Galaxy classification into groups of similar morphological appearances facilitates the study of the associated stages of development. Classifying the enormous databases of galaxies would allow scientists to test their theories about the physical process governing star-formation and galaxy evolution to reach solid conclusions. However, the process of galaxy classification proved to be challenging (González et al. 2018). Traditionally galaxies were classified by reliable human experts. The development of advanced telescopes and imaging techniques resulted in extremely large databases of galaxy images such as the Sloan Digital Sky Survey (SDSS) (Conselice et al. 2016). The enormous amount of galaxy images has rendered manual expert classification impossible. Moreover, the far distances of newly discovered galaxies caused their images to be of low quality, which further complicated the manual classification process. Additionally, numerous galaxies have complicated natures that affect the accuracy of their manual classification (Ellison et al. 2013). Accordingly, astrophysicists consider the classification of detected galaxies as a long-term goal. This gave momentum to the

a b

Fig. 1: (a) Edge-on galaxy NGC 891 and (b) face-on galaxy Messier 74.

development of automated galaxy classification techniques (Cecotti 2020).

Recently, numerous studies investigated the development of automated computational techniques to classify galaxies as well as analyzed their morphology. Several classification techniques based on machine learning were developed to facilitate automated galaxy classification (Abd. Elfattah et al. 2014). During the last decade, research concerning edge-on galaxy classification has accelerated. This reflects the growing interest in understanding the evolution process of edge-on galaxies to explain observed properties of our Milky Way as an edge-on galaxy (Kautsch et al. 2006).

The problems facing edge-on galaxy classification are primarily due to their low brightness and background stars, galaxy center shift from the center of the image as well as slanted galaxy axes within their images. For this purpose, the new technique introduces an adaptive brightness threshold for galaxy pixel detection and proposes a mathematical treatment of the detected galaxy to reorient its axis so that it would be vertical and coincide with the y-axis. Finally, the novel technique puts forth a new galaxy slimness weighting factor along with its threshold that would identify edge-on galaxies from face-on galaxies.

2 RELATED WORKS

Different techniques are applied for the automated classification of edge-on galaxies along with other galaxy morphologies. Some techniques extract features from galaxy images and use them for galaxy classification. Other techniques employ different machine learning approaches. However, classification accuracies are the main challenge for different automated classification techniques.

Dhami et al. (2017) proposed a knowledge-based approach for the classification of edge-on galaxies. The approach relies on using galaxy binary images and finds ratios representing galaxy width divided by galaxy height, measured in x and y directions respectively. Then, the image was rotated 45 degrees and the ratio was evaluated again. If the value of either ratio was higher than a set threshold, the galaxy was considered an edge-on. However, classification accuracy is not given.

Abd El Aziz et al. (2017) presented an automatic detection technique for galaxy morphology based on image-retrieval. This technique detects galaxy type within an image as well as most similar images. It was tested on galaxies from the EFIGI catalog (Baillard et al. 2011) and achieved an accuracy of 97.5% on edge-on galaxy classification.

Shamir (2009) proposed a supervised learning algorithm that can classify galaxies automatically from their images. The algorithm was trained with manually classified galaxies. A set of image features was extracted from galaxy images, and Fisher scores (Bishop 2006) were employed to select the most informative features. Finally, test images were classified using a simple Weighted Nearest Neighbor rule. The algorithm was tested on galaxy images from the Galaxy Zoo catalog (Lintott et al. 2011), and achieved an accuracy of 90% on edge-on galaxy classification.

González et al. (2018) presented an automatic galaxy detection and classification method, based on a novel data augmentation procedure. The models were trained employing deep learning techniques and convolutional neural networks (CNNs). The detection and classification methods were trained utilizing different datasets. The model achieved about 78% accuracy on the classification of edge-on galaxies.

Domínguez Sánchez et al. (2018) presented a galaxy classification scheme. The classifications were obtained with Deep Learning algorithms using CNNs. After being

trained with the Galaxy Zoo 2 catalog (Willett et al. 2013), the scheme achieved an accuracy of 97% on edge-on galaxy classification.

3 PROPOSED METHODOLOGY

A novel technique for automated classification of edge-on galaxies based on their unique slim shape with its high aspect ratio has been developed. This technique is mainly designed to identify edge-on galaxies from face-on galaxies. The novel technique directly employs galaxy brightness data. In the next sections, the novel edge-on galaxy classification technique is presented.

The novel technique consists of four different steps. The first step is galaxy detection in which pixels representing the galaxy are identified as those of the largest spot in the image, after application of a novel adaptive brightness threshold. The second step is galaxy line fitting. In this step, the equation of the line that best fits the galaxy axis is identified. The third step is galaxy reorientation in which the galaxy is reoriented so that its axis would coincide with the y-axis of the image. The fourth and last step is galaxy classification. In this step, a novel slimness weighting factor is evaluated and compared to its slimness threshold to identify edge-on galaxies from face-on galaxies.

3.1 Galaxy Detection

The galaxy detection module is applied to the brightness data of galaxy images. Due to the variation of galaxies and image brightnesses, an adaptive brightness threshold for galaxy pixel identification is employed. For the current work, adaptive brightness threshold is evaluated as a function of the image brightness mean value multiplied by a factor of proportionality (N), Equation (1). This adaptive brightness threshold equation is meant to accommodate variations of image brightness while taking into account the small number of pixels representing edge-on galaxies in comparison to other types of galaxies. The mean value of brightness is evaluated by dividing the sum of image pixel brightness by the total number of pixels in the image, Equation (2). The factor of proportionality (N) value is optimized for higher edge-on galaxy classification accuracy in a later stage.

$$B_{at} = N * B_m \tag{1}$$

$$B_m = \frac{1}{n_p} \sum_{p=1}^{n_p} B_p \tag{2}$$

where B_{at} is adaptive brightness threshold, N is factor of proportionality, B_m is mean brightness of image pixels, n_p is number of pixels in the image, p is pixel number and B_p is brightness of pixel number p.

The adaptive brightness threshold B_{at} is employed to convert the greyscale image of the galaxy into a binary image. Then the Matlab ® function "bwlabel" is applied to detect and label connected white spots in the image. The largest spot, considered to represent the galaxy, is detected as the one labeled with the higher frequency label. The image matrix columns and rows of the largest spot's points are identified as x and y locations of the pixels in the galaxy image respectively.

3.2 Galaxy Line Fitting

For identifying galaxies' main axis, the least squares fitting technique is employed on the orthogonal distances between detected galaxy pixels x, y locations and the fitting line. For this purpose, Matlab ® function "linortfit2" is utilized. The "linortfit2" function returns the intersection point of the orthogonally fit line with the y-axis as well as its slope, Equation (3). This fitting line is considered the galaxy axis. The axis is employed to standardize the orientation of the galaxy before evaluating the galaxy slimness weighting factor.

$$y = mx + b \tag{3}$$

where m is slope of the orthogonally fit line, b is intersection point of orthogonally fit line with the y-axis, and x, y are x, y locations of orthogonally fit line's pixels.

In very rare cases, the orthogonally fit line would be perfectly parallel to the y-axis ($m = \pm\infty$ and $b = \pm\infty$), due to the fit line inclination angle being equal to $\pm\pi/2$. In this case, a data conditioning step is carried out, in which all initial locations of the galaxy pixels are rotated 45 degrees ($\pi/4$) around the origin point $(0, 0)$ according to Equations (4) and (5). Then, the line fitting process is repeated using the conditioned data that replace the initial locations of galaxy pixels for all following processes. This conditioning allows the galaxy data to accept the line fitting process and the rest of the processes, without affecting the outcome of the classification process.

$$y_{p0c} = x_{po}\sin\left(\frac{\pi}{4}\right) + y_{p0}\cos\left(\frac{\pi}{4}\right) \tag{4}$$

$$x_{p0c} = x_{po}\cos\left(\frac{\pi}{4}\right) - y_{p0}\sin\left(\frac{\pi}{4}\right) \tag{5}$$

where x_{p0c}, y_{p0c} are x, y conditioned initial location of detected galaxy pixel p and x_{p0}, y_{p0} are x, y initial location of detected galaxy pixel p.

3.3 Galaxy Reorientation

Edge-on galaxy shapes are quite slim. However, a weighting factor has been considered required for the computer to automatically identify edge-on galaxies. Orienting the galaxy vertically or horizontally has been

considered the most feasible way to identify the ratio of its length to its width. For this purpose, galaxies being classified are reoriented vertically.

The first step in galaxy reorientation is to shift the galaxy y location so that the galaxy axis represented by the orthogonally fit line would intersect with the y-axis at the origin according to Equation (6). During this step, x locations of galaxy pixels are not affected, Equation (7). The second step is to rotate the pixels of the galaxy by the angle (θ) counterclockwise (CCW) required for its axis, represented by the orthogonally fit line, to coincide with the y-axis according to Equations (8) and (9). The rotation angle (θ) is derived from the equation of the orthogonally fit line, Equations (10) and (11).

$$y_{p1} = y_{p0} - b \tag{6}$$

$$x_{p1} = x_{p0} \tag{7}$$

$$y_{p2} = x_{p1}\sin(\theta) + y_{p1}\cos(\theta) \tag{8}$$

$$x_{p2} = x_{p1}\cos(\theta) - y_{p1}\sin(\theta) \tag{9}$$

$$\theta = \frac{\pi}{2} - \varphi \tag{10}$$

$$\varphi = \tan^{-1}(m) \tag{11}$$

where x_{p0}, y_{p0} are x, y initial location of detected galaxy pixel p, x_{p1}, y_{p1} are x, y location of shifted galaxy pixel p, x_{p2}, y_{p2} are x, y final location of galaxy pixel p, φ is slope angle of fitting line and θ is reorientation angle.

3.4 Galaxy Classification

Preliminary runs found that the ratio between galaxy width and its length would not efficiently identify edge-on galaxies from the others. A higher measure for weighting was considered a requirement. For this purpose, galaxies are reoriented to get their axes coinciding with the y-axis so that squared x locations would further weight the identification of galaxy slimness, in a fashion close to the least squares method. The ratio between reoriented galaxy average of pixels x squared and the galaxy length (L_y) in pixels, as expressed in Equation (12), has been found to represent a feasible slimness weighting factor (R_{xy}). L_y is evaluated as the difference between maximum and minimum values of y_{p2} of galaxy pixels.

Edge-on galaxies would exhibit R_{xy} values lower than or equal to a specific R_{xy} threshold (R_{xyt}) while face-on galaxies would experience higher values than the threshold. The value of R_{xyt} threshold is optimized for higher classification accuracy in a later stage.

$$R_{xy} = \frac{1}{n_g\, L_y} \sum_{p=1}^{n_g} x_{p2}^2 \tag{12}$$

where R_{xy} is slimness weighting factor, n_g is number of galaxy pixels and L_y is galaxy length in pixels.

4 EXPERIMENTAL RESULTS AND DISCUSSION

The proposed technique was developed using Matlab ® version 2017 running on a DELL Inspiron laptop with 8 GB of RAM and a 64 Bit core i7-36120QM processor. The laptop employed the Windows 10 operating system. In the next sections, experimental results are demonstrated along with a sample of runs.

4.1 Galaxy Dataset

The current edge-on classification technique has been optimized for the EFIGI catalog (Baillard et al. 2011). For this purpose, it is trained using a set of 1800 galaxies from the EFIGI catalog; 800 of them are edge-on galaxies and 1000 of them are face-on galaxies. This collection includes problematic faint galaxy images with high brightness backgrounds and bright stars to ensure the robustness of the current technique. Finally, the developed technique is tested on the full set of galaxies of the EFIGI catalog (4458 galaxies) including all its problematic galaxies.

4.2 Galaxy Detection

An image of edge-on galaxy IC 210, displayed in Figure 2(a), is employed as a demonstration sample for the developed technique. Figure 2(b) shows the galaxy after being converted to a binary image using its own adaptive brightness threshold, taking the factor of proportionality (N) equal to two. In the later stage, N values are optimized for higher classification accuracy. The developed galaxy's binary image is treated with Matlab ® function "bwlabel". The "bwlabel" function labels spots of eight or more connected white pixels in the image with different numbers. In the current work, the spot with the largest number of connected points is considered the area representing the galaxy and is isolated, as depicted in Figure 2(c).

4.3 Galaxy Line Fitting and Reorientation

The first step required for line fitting is to identify locations of pixels x, y of the galaxy being classified. These pixel x, y locations, plotted in Figure 3, are employed for the rest of the process. The x, y locations are utilized along with the least squares fitting technique, using Matlab ® function "linortfit2", to identify the galaxy's orthogonally fit line, considered as its axis, as displayed in Figure 4. Current results indicate that the orientation angle for galaxy IC 210 within its image is $-25°9'40''$. The results are verified manually and found to be perceptibly accurate.

A galaxy's orthogonally fit line is reoriented so that it coincides with the y-axis to facilitate employing the weighting factor. Figure 5 shows the galaxy after shifting vertically to move the orthogonally fit line intersection

a b c

Fig. 2: Galaxy IC 210, (a) original galaxy image, (b) after applying adaptive brightness threshold and (c) after isolating the galaxy area.

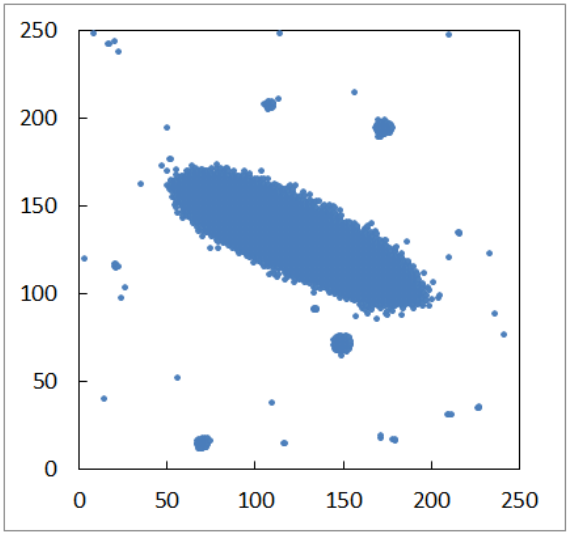

Fig. 3: Plotted x, y values of galaxy IC 210 pixels.

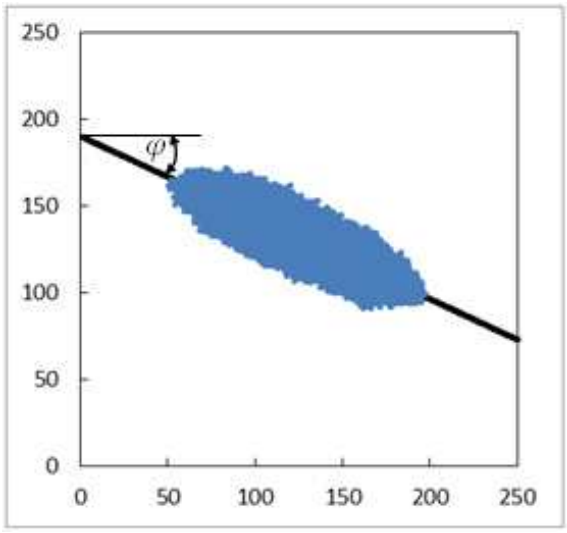

Fig. 4: Galaxy IC 210 plotted along with its orthogonal fitting line.

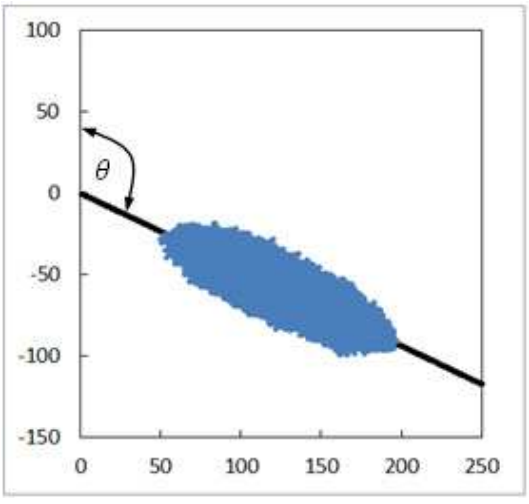

Fig. 5: Plot of galaxy IC 210 after being shifted vertically.

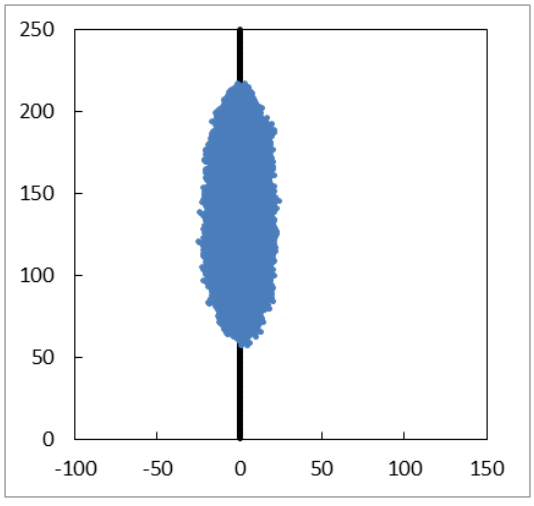

Fig. 6: Plot of galaxy IC 210 after being rotated θ around the origin.

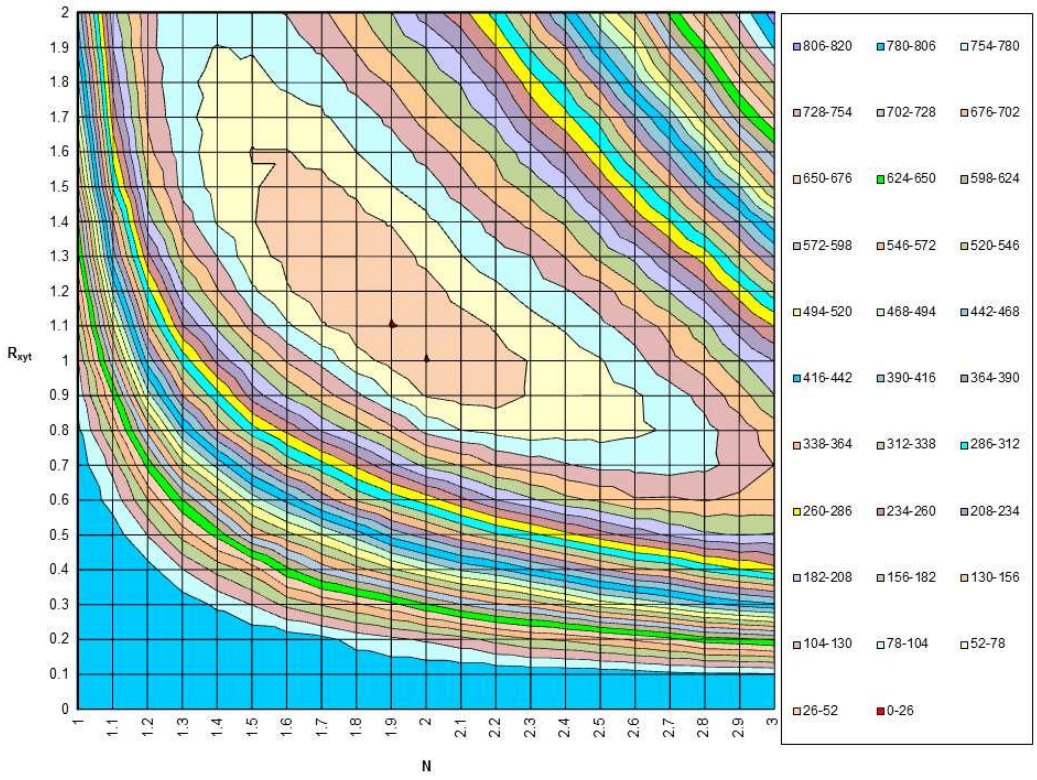

Fig. 7: Number of wrongly classified galaxies out of the 1800 EFIGI catalog sample at different N and R_{xyt} threshold values.

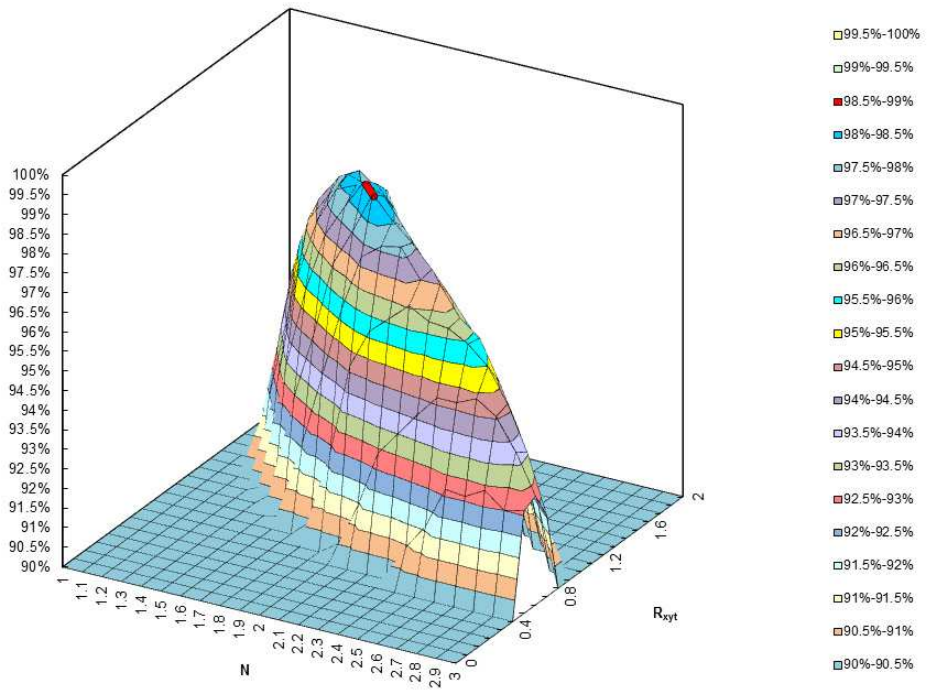

Fig. 8: Percentage of successfully classified galaxies out of the 1800 EFIGI catalog sample at different N and R_{xyt} threshold values.

Table 1: Results of the Training Set

Galaxy type	No. of galaxies	True Classified	False Classified	Accuracy
Edge-on	800	786	14	98.2%
Face-on	1000	989	11	98.9%
Total	1800	1775	25	98.6%

Table 2: Summary of the Developed Technique Results

Galaxy type	No. of galaxies	True Classified	False Classified	Accuracy
Edge-on	1054	1012	42	96.0%
Face-on	3404	3337	67	98.0%
Total	4458	4349	109	97.5%

Table 3: Comparative Results for Galaxy Classifications

Related work	Year	Employed technique	Accuracy
Shamir	2009	supervised learning algorithm using Fisher scores	90 %
Abd el aziz et al.	2017	image retrieval approach	97.5%
Sánchez et al.	2018	deep learning algorithm using CNNs	97 %
Gonzalez et al.	2018	deep learning and data augmentation	78%
Proposed technique	2021	mathematical treatment of brightness data	97.5%

point with axes to the point of origin. Figure 6 displays the shifted galaxy after rotation with calculated angle ($\theta = 115°9'40''$) CCW around the point of origin, resulting in the orthogonally fit overlapping the y-axis.

4.4 Galaxy Classification

In this section, the process of optimizing R_{xyt} and N factors is outlined and the final edge-on classification results achieved by the current study are presented along with a comparison with reviewed results. Finally, the processing load of the current technique is investigated using processing time on an average laptop as a performance measure.

4.4.1 Optimization of R_{xyt} and N factors

An investigation was conducted to evaluate the optimum values of the proportionality factor (N) and its coupled slimness threshold (R_{xyt}) for the studied EFIGI catalog. In this investigation, a training set comprised of 800 edge-on galaxies and 1000 face-on galaxies from the EFIGI catalog was employed. Investigated N values ranged from 1 to 3 and investigated R_{xyt} values ranged from 0 to 2. The total numbers of wrongly classified galaxies and achieved accuracies at different N and R_{xyt} combinations were evaluated and are presented in Figure 7 and Figure 8 respectively. The results affirm that the N value of 2 along with the R_{xyt} value of 1 achieve the best classification accuracy. The figures demonstrate that high accuracy is achievable at a wide range of N and R_{xyt} value combinations. The achieved overall accuracy of the training set is 98.6%, with 98.2% accuracy in classification of edge-on galaxies and 98.9% accuracy in classification of face-on galaxies, see Table 1.

4.4.2 Accuracy of the developed technique

Finally, the classification accuracy of the developed technique has been tested utilizing the full EFIGI catalog, 4458 galaxies. In this test, the developed technique achieved an overall accuracy of 97.5% with 96% accuracy in the classification of edge-on galaxies and 98% accuracy in the classification of face-on galaxies, as displayed in Table 2. Accuracy results have not significantly varied from the training set results which demonstrate low sensitivity of the technique to N and R_{xyt} values.

A comparison between accuracies of current work and reviewed work is presented in Table 3. The comparison shows the high capacity of the current novel technique to classify edge-on galaxies, which would allow it to be soundly employed for the required classification process. It also demonstrates that the current technique would have the potential as a basis for further development.

4.4.3 Processing load of the developed technique

To demonstrate the capacity of the developed technique to process galaxies in a relatively short time, processing times of all tested galaxies are evaluated on a DELL Inspiron laptop with 8 GB RAM and a 64 Bit core i7-36120QM processor. The histogram of tested edge-on galaxy processing times is displayed in Figure 9(A), while the histogram of tested face-on galaxy processing times is depicted in Figure 9(B). The histogram of all tested galaxy processing times is featured in Figure 9(C). The total processing time of the 4458 galaxies is about 19 minutes,

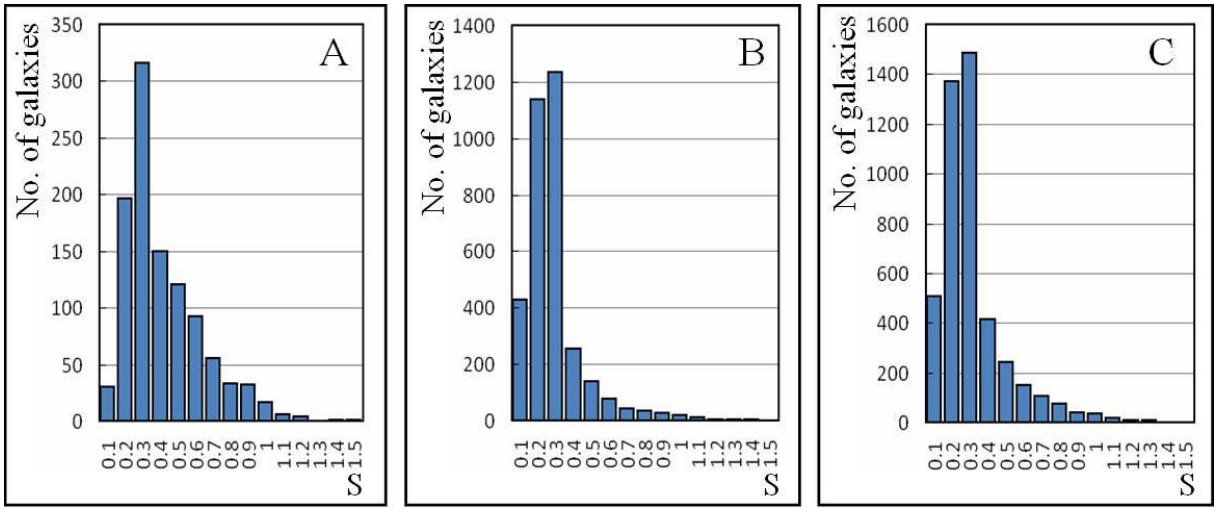

Fig. 9: Histograms of processing time of galaxies, (A) tested edge-on galaxies, (B) tested face-on galaxies and (C) all tested galaxies.

about 0.26 seconds per galaxy. This fast processing time allows for the classification of large catalogs without the need for advanced computing facilities.

5 CONCLUSIONS

In the current work, a novel classification technique that differentiates between edge-on galaxies and face-on galaxies has been developed. It relies on mathematical treatment of galaxy brightness data by introducing adaptive brightness threshold for galaxy pixel identification, slimness weighting factor and slimness weighting threshold for edge-on galaxy classification. The technique has been optimized using a training set of 1800 galaxies from the EFIGI catalog. The optimized technique was employed to classify the full EFIGI catalog and achieved an overall classification accuracy of 97.5%. The processing time of the current technique on an average laptop is about 0.26 seconds per galaxy. Accordingly, the current technique has the capacity to be soundly employed for the classification of larger catalogs without the need for special computing facilities.

References

Abd El Aziz, M., Selim, I. M., & Xiong, S. 2017, Scientific Reports, 7, 4463

Baillard, A., Bertin, E., de Lapparent, V., et al. 2011, A&A, 532, A74

Bishop, C. 2006, Pattern Recognition and Machine Learning (Springer)

Cecotti, H. 2020, International Journal of Machine Learning and Cybernetics, 11, 1839

Conselice, C. J., Wilkinson, A., Duncan, K., & Mortlock, A. 2016, ApJ, 830, 83

Dhami, D. S., Leake, D., & Natarajan, S. 2017, in AAAI Conference, 719, DOI: 10.13140/RG.2.2.29961.65125

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., & Fischer, J. L. 2018, MNRAS, 476, 3661

Elfattah, M. A., El-Bendary, N., Elsoud, M. A. A., et al. 2014, Advances in Intelligent Systems and Computing, Springer, 237, DOI: 10.1007/978-3-319-01781-5_21

Ellison, S. L., Mendel, J. T., Scudder, J. M., Patton, D. R., & Palmer, M. J. D. 2013, MNRAS, 430, 3128

Elsayed Abd Elaziz, M., Hosny, K. M., & Selim, I. M. 2019, Soft Computing, 23, 9573

González, R. E., Muñoz, R. P., & Hernández, C. A. 2018, Astronomy and Computing, 25, 103

Gott, J. Richard, I., Jurić, M., Schlegel, D., et al. 2005, ApJ, 624, 463

Kautsch, S. J., Grebel, E. K., Barazza, F. D., & Gallagher, J. S., I. 2006, A&A, 445, 765

Khalifa, N. E. M., Taha, M. H. N., Hassanien, A. E., & Selim, I. M., 2018, in IEEE conference, 978, DOI: 10.1109/ICCSE1.2018.8374210

Lintott, C., Schawinski, K., Bamford, S., et al. 2011, MNRAS, 410, 166

Shamir, L. 2009, MNRAS, 399, 1367

Sparke, L. S., & Gallagher, John S., I. 2007, Galaxies in the Universe: An Introduction

Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835