Galaxy Classification Using Optimal Convolution Neural Network

Dr. Goutam Sarker, Associate Professor,
Senior Member IEEE
Computer Science and Engineering Departmennt
National Institute of Technology, Durgapur, India
goutam.sarker@cse.nitdgp.ac.in

Saharsh Ananta Jaiswal
B.Tech Student
Computer Science and Engineering , Department
National Institute of Technology, Durgapur, India
saj18u10222@btech.nitdgp.ac.in

Abstract – **A galaxy classification system using Optimal Convolution Neural Network has been designed and developed. Broadly there are five different types of galaxies. They are elliptical, spiral, disk, lenticular and irregular. We have proposed an optimal deep convolution neural network which learns the features of the different types of galaxies in the form of a set of filters and after the learning is over recognizes the type of the unknown galaxies. The CNN uses stratified data set of 300 images of different galaxies. With this, the optimal point with data size (150) as well as number of filters (130) of neither overshooting nor undershooting is found out to get the optimal network complexity – which highly improves the performance evaluation. The performance evaluation of the CNN architecture used for galaxy classification in terms of its accuracy, precision, recall and f-score is quite satisfactory. Also the training and testing time is affordable.**

Keywords—Deep learning, Convolutional Neural Network, Image Classification, Galaxy, Types of Galaxies, Performance Evaluation, Holdout Method, Confusion Matrix, accuracy, Precision, Recall, f-score

I. Introduction

Through radio telescopes and / or other optical instruments, we can get photometric data for hundreds of millions to billions of stars and galaxies [1,2,3]. Due to this huge voluminous data, it is quite impossible for human beings or even astronomy experts to manually classify them.

Previously galaxies classifications were done through visual inspection of two-dimensional images of galaxies and appearance based classification. Even if we assume that accurate and correct classification of abnormally huge quantity of data done by experts of astronomy is far more reliable, it is too much time consuming which is totally unaffordable. Thus there is a tremendous need to separate those images of terrestrial bodies like the different types of galaxies by an automatic procedure especially a neural network based machine learning approach [14,15,16,17,18,19] like a deep convolution neural network.

II. ASTRONOMY AND GALAXIES

Only approximately two to three thousand stars are visible by a naked eye. Merely trillions of stars and terrestrial bodies are visible in case we take the aid of most powerful telescope. A Galaxy [2,3,4] is a cluster or a group of billions of stars gathers together. "Milky Way" is an example of such galaxy. In this galaxy, the sun along with its planets is housed. One end to the other of a galaxy may be as far as trillions of light year – which is abnormally large and unimaginable. Large number of huge clouds of dusts and gaseous elements in galaxies are called Nebulas. Our galaxy "Milky Way" and the nearest galaxy "Andromeda" is fas away by 20 billion light year.

The combination or the cluster of a huge number of galaxies is called Super Galaxy. Our galaxy "Milky Way", "Andromeda" along with such other approximately 24 other galaxies form one super galaxy. "Hydra" is a super galaxy with several thousand galaxies housed inside it.

Galaxies are conventionally classified [2,3,4] according to either their ***(1) Shape or (2) Size***

(1) ***Shape:*** We can classify galaxies by their shape or in the way they appear to us. This is termed as Hubble Classification. Some prominent views are given below

i. Elliptical Galaxies
ii. Spiral Galaxies
iii. Disk

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

iv. Irregular
v. Lenticular

(2) ***Size:*** According to this, galaxies range from dwarfs (less than one million stars within it, diameter only a few light year) to super giants (with over a trillion stars, with diameter over 6×10^5).

Galaxies may either be isolated (Singleton) or form either a small cluster or a large cluster (one such example is Virgo Cluster). Galaxies may also form a cluster of clusters – called Super Cluster.[2]

III. Over fitting and Under fitting problem in ConvolutioN Neural Network

Any Convolution Neural Network (CNN) [7,8,9,10,11,12,13] has three layers (i) convolution, (ii) pooling and (iii) fully connected (third) layer. [9,10,11,12,13]. The convolution and pooling layers may be repeated to get desired activation map size. When these are combined together, a complete CNN structure is formed as indicated in Fig. 1.

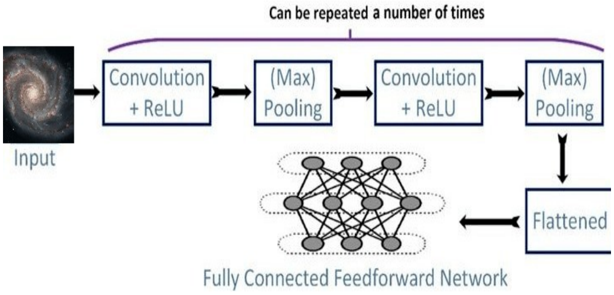


Fig. 1 Present CNN model for galaxy classification

Like any other ANN model, there might be some problems due to either over fitting or under fitting in any CNN model.

A. Over Fitting in CNN

This problem is likely to occurs when the training data size is small compared to the network complexity, such that the network is able to "memorize" all the given data instead of "learning" them through optimum generalization.

B. Under Fitting in CNN

This problem is likely to occur when the training data size is large enough compared to the network complexity, such that network "maximally generalizes" without performing effective or fruitful learning.

In a particular CNN model architecture all other network parameters while kept constant, it is only the number of filters at each convolution layer, which govern the complexity of the model.

IV. ALGORITHMS FOR PREPROCESSING AND LEARNING

Pre-processing of data: First step is to pre-process the benchmark data set. This includes noise removal, data augmentation through rotation, translation, flipping.

CNN Learning of the data: Second step is the learning of the benchmark data set. Here we vary both the number of filters as well as the size of the training data set together and each time compute the accuracy with validation data.

Initially we start from a very small data size and huge number of filters. This makes the network complexity too large for the given data set. Here the network is maximally over-shooted.
We calculate the performance evaluation of the system through validation (test) data.

We continuously step by step increase training data size and at the same time decrease the number of filters. Each time as before, we calculate the performance evaluation of the system.

In this way, we come to the end point when all the total amount of data has been used as training data and the corresponding minimum number filters. Similarly we calculate the accuracy of the system.

Now we plot (data size, no. of filters) vs. accuracy. We find the optimal point where the accuracy is maximum for the test data set.

This point is the desired optimal point where we get the desired number of data size and the corresponding number of filters.

A. Galaxy Image Pre processing Algorithm

Input: Total 300 Raw Image of each of size 300*300

Output: 300 Pre-processed Image

Steps:

1. Crop the images to 180x180 pixels, retaining color to removes the noise in the outer part of the galaxies.

2. Perform random data augmentation of the training images data through Rotation, Translation, Flipping.

3. Scale the image by dividing each element by 255.

4. Save the 300 pre-processed image for training and validation

[1]*B. Learning Algorithm*

Input: (1) Pre-processed stratified image data set of size 300. (2) Keras CNN model for training

Output: Trained Keras Model for Galaxy Classification.

Steps:

1. Divide the dataset into train/test split of 2:1 according to Holdout Method when it is given as input for each training and validation

2. Vary both the number of filters as well as the size of the training data set together. Initially start from a very small training data size and huge number of filters.

3. Step by step increase training data size and at the same time decrease the number of filters. Each time we calculate the performance evaluation of the system.

4. Come to the end point when all the total amount of data has been used up as training/ validation data and the corresponding minimum number filters.

Plot (data size, no. of filters) vs. accuracy. Find the optimal point where the accuracy is maximum for the test data set.

V. RESULTS

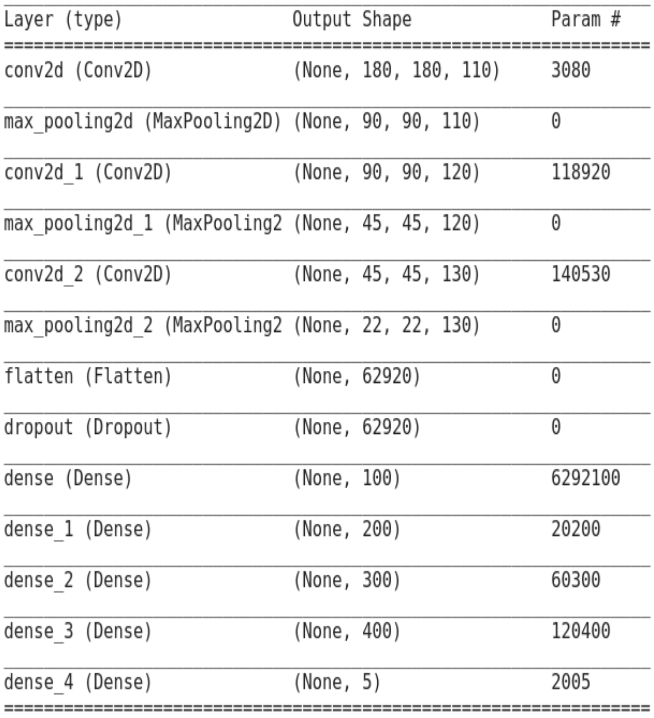
A. Platform Used

Implemented in Python.
Framework used – Tensorflow
Editor/IDE – Google Colab/Jupyter Notebook
Librarires used – Keras-CNN, Numoy, Pandas, Sklearn, Seaborn, Matplotlib etc.
Specifications Used- 12GB Ram/ Nvidia T4 GPU

B. Benchmark Dataset Used

Link for the dataset - https://drive.google.com/file/d/1x3Y4e5j6vI_-uu91Vk-xR-1WlknZquUK/view?usp=sharing

C. Model Summary



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 180, 180, 110)     3080
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 90, 90, 110)       0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 90, 90, 120)       118920
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 45, 45, 120)       0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 45, 45, 130)       140530
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 22, 22, 130)       0
_________________________________________________________________
flatten (Flatten)            (None, 62920)             0
_________________________________________________________________
dropout (Dropout)            (None, 62920)             0
_________________________________________________________________
dense (Dense)                (None, 100)               6292100
_________________________________________________________________
dense_1 (Dense)              (None, 200)               20200
_________________________________________________________________
dense_2 (Dense)              (None, 300)               60300
_________________________________________________________________
dense_3 (Dense)              (None, 400)               120400
_________________________________________________________________
dense_4 (Dense)              (None, 5)                 2005
=================================================================
```

Fig. 2 Model Summary

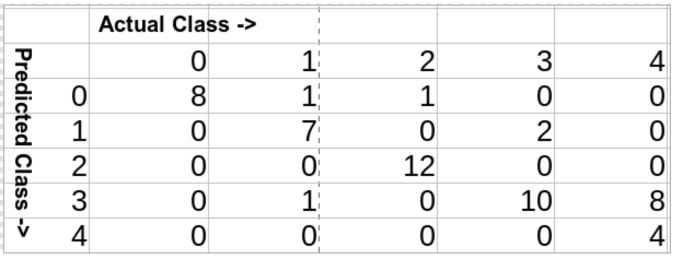
D. Training and Testing Results

There are total 12 steps and all were run for 200 epochs.

No of Starting Filters and Dataset:
Filters – 240
Dataset – 25

No of Ending Filters and Dataset:
Filters – 20
Dataset – 300

The confusion matrix and the accuracy results are shown in a tabularized form below in table 1 and table 2 respectively.

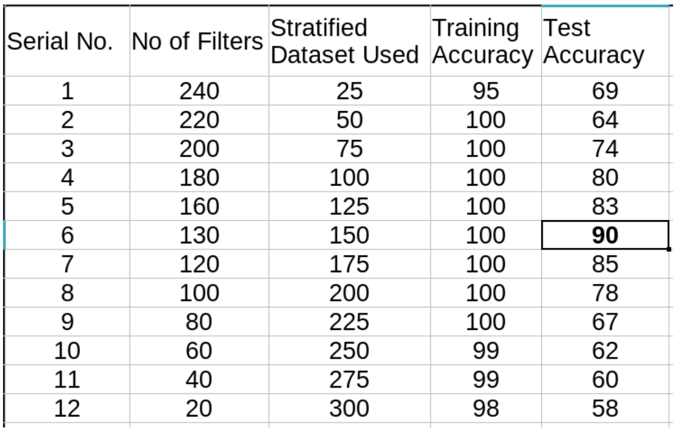


		Actual Class ->				
		0	1	2	3	4
Predicted Class ->	0	8	1	1	0	0
	1	0	7	0	2	0
	2	0	0	12	0	0
	3	0	1	0	10	8
	4	0	0	0	0	4

Table-1 Confusion Matrix

In the above table confusion the numbers 0-4 has the following interpretations:

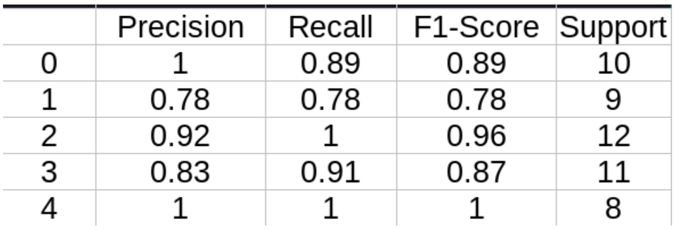
0 - Disk Galaxy
1- Elliptical Galaxy
2 - Irregular Galaxy
3 - Lenticular Galaxy
4 - Spiral Galaxy



Serial No.	No of Filters	Stratified Dataset Used	Training Accuracy	Test Accuracy
1	240	25	95	69
2	220	50	100	64
3	200	75	100	74
4	180	100	100	80
5	160	125	100	83
6	130	150	100	**90**
7	120	175	100	85
8	100	200	100	78
9	80	225	100	67
10	60	250	99	62
11	40	275	99	60
12	20	300	98	58

Table-2 Results indicating no of filters corresponding Training and Testing Accuracy

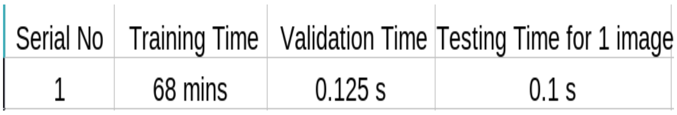
The classification report is presented below in table-3



	Precision	Recall	F1-Score	Support
0	1	0.89	0.89	10
1	0.78	0.78	0.78	9
2	0.92	1	0.96	12
3	0.83	0.91	0.87	11
4	1	1	1	8

Table-3 Precision, Recall, F1-Score and Support

Estimated time taken to run the model in the optimal point (130 filters in each layer and 150 Dataset Size) for 200 epochs was 1 hour and it is presented below in table-4. 33% Train/Test Split in Dataset.



Serial No	Training Time	Validation Time	Testing Time for 1 image
1	68 mins	0.125 s	0.1 s

Table-4 Training & Testing Time Taken at Optimal Point

E. Optimal Point-

In the sixth step, with 150 stratified data, accuracy of **90%** was reached with 130 filters in the convolutional layer.

F. Several Plots

1. Plot of No of Filters/Stratified Dataset vs Step.

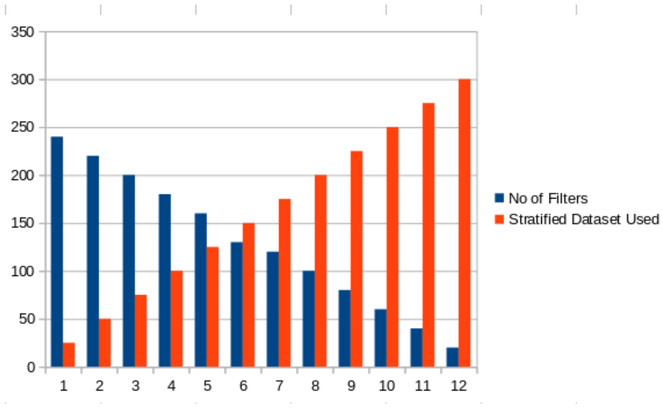
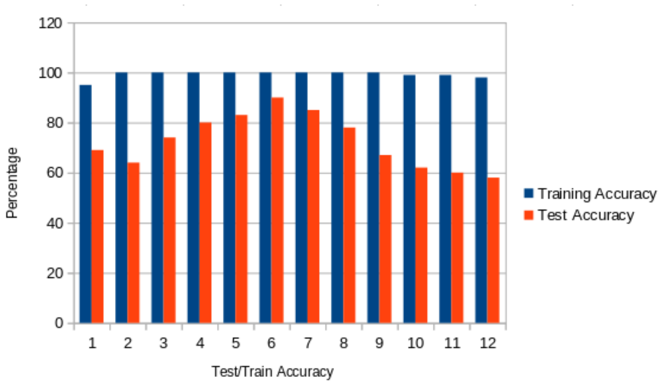


Figure-3 No of Filters/ Stratified Dataset vs Step.

2. Plot of Train/Test Accuracy



.

Figure-4 Training/Testing Accuracy vs Step

3. Optimal Point for Testing Accuracy

The model reaches an optimal point at step no 6. far from the point of overshooting and undershooting, with 130 filters in each convolutional layer and 150 stratified dataset.
It achieved an accuracy of 90% when run for 200 epochs.
It is presented in the figure-6 below

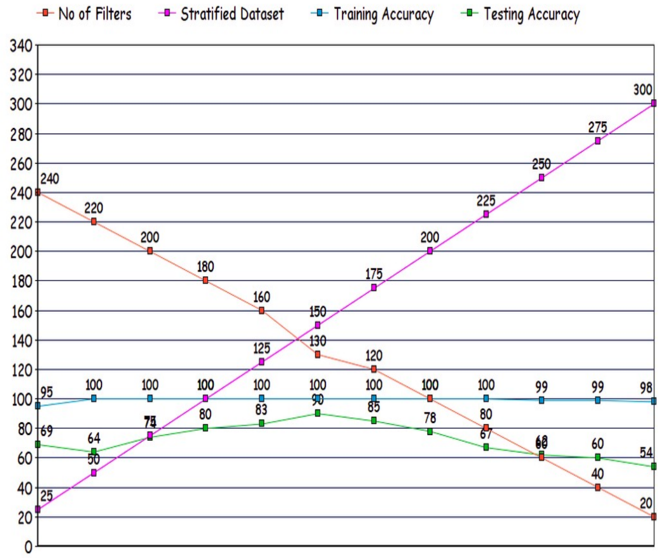


Figure-5 Optimal Point

VI Filters and Feature maps Visualization

Some examples of filters at Convolution 1

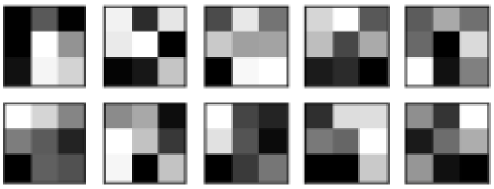


Figure-6 Convolution 1 (110 such filters 3x3)

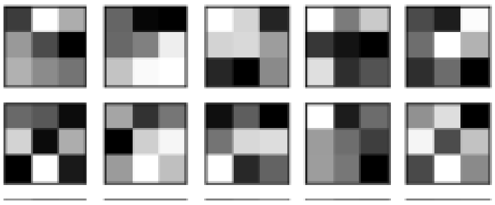


Figure-7 Convolution 2(120 such Filters 3x3)

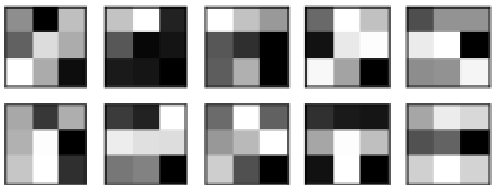


Figure-8 Convolution 3 (130 such Filters 3x3)

Figure-10,11,12 shows the filter weights learned on every convolutional layer. As expected, the first layer captures everything in the image and detects the different galaxy edges and corners etc, from original pixel, then uses the edges to detect simple shapes in second layer filters, then as we go further deeper it checks very specific properties from the images - rounded disk of the disk galaxy.

Figure-9 Disk Galaxy

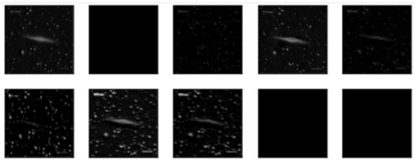


Figure-10 Feature Map Conv1

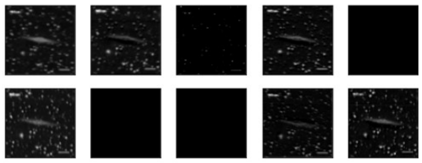


Figure-11 Feature Map Conv2

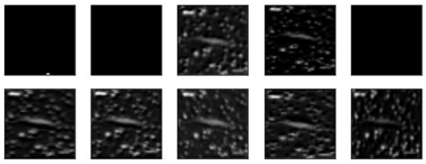


Figure-12 Feature Map Conv3

At the very top layer the filters are very basic and ask very generic questions. Is the galaxy smooth? Rounded? Does it have any special features? As we go further in the layers these questions get more specific. Does it have a bulge in the center? Is there a spiral pattern? .

VII. Conclusion

In the present paper we have proposed and developed a galaxy classification system using optimal deep CNN model. The system finds out on its own, the optimal size of the training benchmark data set and at the same time the optimal number of filters with that training data for the given deep CNN. Thus the system model would be able to find out the exact combination of training data size and number of filters of the CNN such that the system is not prone to either overshooting or undershooting. In that way it enjoys the maximum performance evaluation. The performance evaluation of the present CNN architecture used for galaxy classification in terms of its accuracy, precision, recall and f-score is quite satisfactory. Also the training and testing time is affordable.

.REFERENCES

[1] Stephen Hawking, “A Brief History of Time – From Big Bang to Black Holes”, Bantom Books, 1995

[2] G. Sarker, On the Change in Entropy Due to Accretion and Collision of Black Holes - International Journal of Applied Physics (IJAP), Volume 8 Issue 2, 5-15, May-Aug, 2021, doi:10.14445/23500301/ IJAP-V8I2P102

[3] Baidyanath Basu, “An Introduction to Astrophysics”, PHI, India, 2006

[4] Nikolay Tikhonov1 , Olga Galazutdinova1 , Olga Sholukhova1 , Antoniya Valcheva2 , Petko Nedialkov2 and Olga Merkulo Searching for the brightest stars in galaxies outside the Local Group va3 , RAA 2021 Vol. 21 No. 4, 98(7pp) doi: 10.1088/1674-4527/21/4/98

[5] Nebaaer, C.: Evaluation of convolution neural networks for visual recognition. Neural Networks, IEEE Transactions on 9(4), 685-696 (1998).

[6] Simard, P.Y., Steinkraus, D., Platt, J,C.: Best practices for convolutional neural networks applied to visual document analysis. In: null. P. 958. IEEE (2003)

[7] Zeiler, M.D., Fergus, R.: Visualizing and Understanding Convolution Networks. In: Computer Vision – ECCV 2014, pp. 818-833. Springer (2014)

[8] Sarker, G.(2000),A Learning Expert System for Image Recognition, Journal of The Institution of Engineers (I), Computer Engineering Division.,Vol. 81, 6-15.

[9] Sarker G., Ghosh S. (2020), Biometric Based Unimodal and Multimodal Person Identification with CNN using Optimal Filter Set, - Proceedings of the Global AI Congress 2019, vol 1112. Springer, Singapore

[10] Sarker, G. Some Studies on Convolution Neural Network - International Journal of Comuter Applications (IJCA), Foundations of Computer Science, New York, Vol. 182 - No. 21, DOI 10.5120/ijca201891, pp 13-22, Oct. 2018.

11] Sarker, G. Ghosh S. A Convolution Neural Network for Optical Character Recognition and Subsequent Machine Translation - International Journal of Computer Applications (IJCA), Foundations of Computer Science, New York, Vol. 182 - No. 30, DOI 10.5120/ijca2018918203, pp 23-27, Dec. 2018.

12] Sarker, G. Ghosh S. A Set of Convolution Neural Networks for Person Identification with Different Biometrics - International Journal of Advanced Computational Engineering and Networking, Volume - 7, Issue 6, June - 2019.

[13] Sarker G., A Survey on Convolution Neural Networks - TENCON 2020 IEEEE Region 10 International Conference,pp 923-928 16-19 Nov. 2020, Osaka, Japan

[14] M. Abd Elfattah, N. El-Bendary, M. A. Abu Elsoud, A. E. Hassanien, and M. F. Tolba, An intelligent approach for galaxies images classification, in 13th International Conference on Hybrid Intelligent Systems (HIS 2013), 2013, pp. 167172.

[15] M. Abd Elfattah, N. Elbendary, H. K. Elminir, M. A. Abu El-Soud, and A. E. Hassanien, Galaxies image classification using empirical mode decomposition and machine learning techniques, in 2014 International Conference on Engineering and Technology (ICET), 2014, pp. 15.

[16] J. De La Calleja and O. Fuentes, Machine learning and image analysis for morphological galaxy classification, Mon. Not. R. Astron. Soc., vol. 349, no. 1, pp. 8793, 2004.

[17] M. Marin, L. E. Sucar, J. A. Gonzalez, and R. Diaz, A Hierarchical Model for Morphological Galaxy Classification, in Proceedings of the Twenty- Sixth International Florida Artificial Intelligence Research Society Conference, 2013, pp. 438443.

[18] I. M. Selim, A. E., and B. M.El, Galaxy Image Classification using Non-Negative Matrix Factorization, Int. J. Comput. Appl., vol. 137, no. 5, pp. 48, Mar. 2016.

[19] I. M. Selim and M. Abd El Aziz, Automated morphological classification of galaxies based on a projection gradient nonnegative matrix factorization algorithm, Exp. Astron., vol. 43, no. 2, pp. 131144, Apr. 2017.