Classification of Plants into Families based on Leaf Texture

Zacrada Françoise TREY [1], Bi Tra GOORE[1], K. Olivier BAGUI[2], Marie Solange TIEBRE[3],

[1]Laboratoire de recherché en informatique et télécommunication, INP-HB, BP 1093 Yamoussoukro, Cote d'Ivoire
[2]Laboratoire Instrumentation Image et Spectroscopie, INP-HB, BP 1093 Yamoussoukro, Cote d'Ivoire
[3]Laboratoire de Botanique, Université Félix Houphouët-Boigny, 01 bp V 34 Abidjan 01, Cote d'Ivoire

Summary

Plants are important for humanity. They intervene in several areas of human life: medicine, nutrition, cosmetics, decoration, etc. The large number of varieties of these plants requires an efficient solution to identify them for proper use. The ease of recognition of these plants undoubtedly depends on the classification of these species into family; however, finding the relevant characteristics to achieve better automatic classification is still a huge challenge for researchers in the field. In this paper, we have developed a new automatic plant classification technique based on artificial neural networks. Our model uses leaf texture characteristics as parameters for plant family identification. The results of our model gave a perfect classification of three plant families of the Ivorian flora, with a determination coefficient (R2) of 0.99; an error rate (RMSE) of 1.348e-14, a sensitivity of 84.85%, a specificity of 100%, a precision of 100% and an accuracy (Accuracy) of 100%. The same technique was applied on Flavia: the international basis of plants and showed a perfect identification regression (R2) of 0.98, an error rate (RMSE) of 1.136e-14, a sensitivity of 84.85%, a specificity of 100%, a precision of 100% and a trueness (Accuracy) of 100%. These results show that our technique is efficient and can guide the botanist to establish a model for many plants to avoid identification problems.

Key words:
Classification, Neural Networks, Texture, Plants.

1. Introduction

2.1 Context

Plant classification has been at the heart of several debates among botanists. This is due to the difficulty of clearly defining the different plant families [1]. For decades, the field of botany has remained in its traditional and manual practices, probably due to the lack of information and/or computer skills of its specialists. Thus, for the morphological classification of the plants, a common, flagship and popular activity, botanists in Africa still work with dichotomous keys that are used by visual inspection of the systematist: a botanist, specialized in plant identification [2]. This classification process is slow and difficult to perform. While, with the development of new computer fields such as artificial intelligence, several automatic classification methods have emerged. Artificial intelligence is defined as the set of means, theories, rules and techniques used to create automata of machines, robots capable of simulating human behavior. Machine learning, one of the predictive methods of Artificial Intelligence, can facilitate classification through classifiers such as neural networks. The objective of our work is to propose an effective model for classifying plant species into families. Among other things, such a model will enable botanists to save time by correctly identifying a plant and making the right decision.

In this paper, we propose an automatic plant classification technique based on neural networks; it is a methodical approach, which from images of plant leaves, allows us to decide based on the measurement of leaf texture parameters of plant species of the Ivorian flora. Indeed, several characteristics of the plant, such as shape, color etc... have been combined with the neural network classifier without giving satisfactory results. The global approach used is to start from a well-known set of plant species families; then, check the relevance of our model by submitting the automatic classification of the species of these same families to it.

2.2 State of the Art

VIJAY SATTI et al have developed a computer-efficient method of plant classification using digital image technology. Their approach is based on the extraction of color and shape characteristics because, according to them, the combination of these characteristics can achieve the desired results. The resulting model was tested on a plant basis. Thus, for the Artificial Neural Network (ANN) classifier, the model gave an accuracy of 93.3%, while with the k-nearest neighbor (KNN) classifier, the accuracy is 85.91%. They conclude that neural networks are efficient for plant family determination [4], however, more leaf color and shape features would be required for better results [5] [6].

ARUN PRIYA and Al. [7] understood that plant classification has become an active field of research because most plant species are not only unknown but also often threatened with extinction. They classified the plants in their locality with the Support Vector Machine (SVM) classifier; pre-processing, feature extraction and classification are the steps that allowed them to implement their approach. The pre-processing phase involved typical

Manuscript received February 5, 2021
Manuscript revised February 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.2.24

image processing steps such as grayscale transformation. Their approach was tested on the international Flavia database. They obtained a good classification rate of 94.5%, however the performance of the SVM classifier could be improved to minimize the classification error [8].

To facilitate automatic plant classification, it is essential to transfer the botanists' descriptions of plant characteristics to a computer. For this purpose, the detection of dots on parts of plant leaves is emphasized by authors Shubham Lavania et al. [9]. Plant leaves are scanned, and the resulting images are stored in a computer. One of the performances of advanced multimedia techniques in the implementation of real-world species identification systems is the storage of images and the extraction of their characteristics [10]: Key points are calculated and stored, and then the number of key points is mapped to the database for the leaf recognition. The classification was made by the scalar invariant Fourier transform (SIFT) based on key descriptors. The model was tested on plant species from the Flavia database with an accuracy of 87.5%, a rate that could be improved if better shape features such as slimness and roundness were detected.

In addition to shape and color features, Kadir Abdul and Al [11] include leaf texture. Using the Polar Fourier Transform (PFT), they extract the fineness, roundness, and dispersion of the leaf. By applying probabilistic neural networks to a set of plants in the Flavia (the international basis of plants) database, the model obtains a 93.75% accuracy for the classification of these plants. Most of the morphological characteristics of the leaf are considered in this type of classification; however, studies by Jagadeesh and Al [12] show that the choice of another classifier such as ANN could improve the accuracy rate.

Artificial Neuron Networks (ANNs) is a well understood and well mastered data processing technique. It provides identification and control functionalities and extend the classical techniques of non-linear automation to provide more efficient and robust solutions [13]. This is one of the reasons why Aimen Aakif et Al. [14a] based their classification on artificial neural networks (ANNs). By considering the morphological characteristics leaf, they developed a model for automatic plant classification. The algorithm developed for the cause is composed of three steps, namely pre-processing, extraction of morphological characteristics and classification. The neural network is composed of as many layers as the number of characteristics entered. The transfer function chosen is the sigmoid function. The model was simulated on 817 images from the Flavia database. To evaluate its efficiency, it was tested on local plant base samples such as ICL. The average precision was 96%. This is a much better rate than its predecessors, but we believe that it can be improved if the transfer function is combined with another one such as hyperbolic tangent.

2. Material and Method

2.1 Material

This part concerns the conditions of acquisition and the nature of the equipment used for our study. It comprises two main stages: the first stage, called sample collection, includes the search for information and then harvesting; and the second stage, the acquisition of images. This stage is divided into two parts which are the digitization and storage of the images.

(i) Step 1: Collection of Sample

This step consists in having information on the existing plants in Ivory Coast and the place of their harvest.

- Search for Information

The Global Biodiversity Information Facility (GBIF) is an international network of researchers. It operates through nodes of participation by providing information on when and where living species were discovered. The GBIF.org index contains hundreds of millions of species records. This allows scientists, researchers, etc. to reuse this data in scientific publications. For better visibility, we have extracted the names of ivory coast plants, their genus and family. Thus, to remove redundancies, we wrote a program executed in python to have Excel files. Then, using a pivot table, we created a file that groups together the plant families that will be the subject of our study.

- The Harvest

Based on our file containing the family names of the plants, we proceeded to collect three families of plants of interest for our study. These harvests took place in the forest of the Centre National Center of Floristics (NCF) of Ivory coast. They took place over two weeks, under the supervision of a systematist. Two hours after each harvest, we spread out the plant samples (roots, leaves, stems, flowers, fruits) on a bench. We placed them carefully on newspaper so that they were well covered and flattened, and then we proceeded to identify them using the botanists' identification technique: For each sample, we wrote on a sheet of paper the name of the plant, its genus, its family, and the date of harvest. We affixed this identification on the newsprint corresponding to the sample concerned. All these samples were put in press. Finally, we embalmed them with an insecticide to prevent parasites (termites, ants...) from eating them and we placed them in the NCF herbarium for two days. Each sample having taken a rather flat shape, we proceeded to scan them.

(ii) Step 2: Acquisition of the Images of The Leaves

This step consists of the transformation of the physical sample of the harvested plant into a photographic sample and the storage of this sample.

- Digitization

We place each sample of the same plant on a white A1 size paper. We gently spread the samples in the flat tray of the scanner screen, and we proceed to scan the different parts of the plants, using an EPSON GT 20000, 600 dpi scanner.

- Images storage

A directory is created with a database of images, containing the three families of plants scanned: APOCYNACEAS, RUBIACEAS and EUPHORBIACEAS. We organize them in the following way.

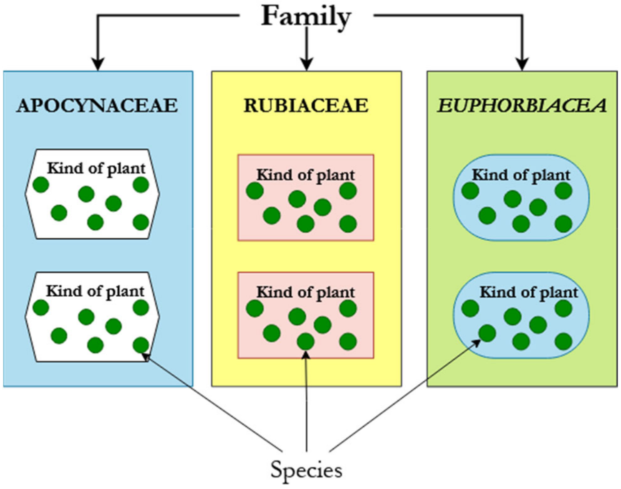

Fig. 1 Organization of image directory

2.1 Method

The technique developed for the classification of our plants consists of four steps: leaf segmentation, feature extraction, resampling and actual classification.

(i) Step 1: Leaf Segmentation

The leaf is the vegetative character of the plant that captures the light energy. It is of interest for our study insofar as it will allow us to highlight the characteristics of its texture. This will be done by cutting out the part of the image of the leaf thanks to well defined contours. In the attribute space, the two-dimensional observations thus obtained constitute two quite distinct classes: that relating to the sheet and that relating to the background of the image. To make the segmentation of the sheet perfect, we proceeded to the conversion of the color image into gray level by applying the equation below [14]:

$$I_g = 0.2989 \times R + 0.5870 \times G + 0.1140 \times B \qquad (1)$$

Where:
- *Ig*: grayscale image
- *A*: Red component of the color image
- *G*: green component of the color image
- *B*: blue component of the color image

The grayscale Ig image is reconverted to binary using the global thresholding technique by the Ots'u method. This method is used to perform automatic thresholding based on the shape of the image histogram. The algorithm then assumes that the binary image contains only two classes of pixels, (i.e., foreground and background) and calculates then the optimal threshold that separates these two classes so that their intra-class variance is minimal [15].

This binary image will be used as a mask for the extraction of the pixel intensity of the region defined by the leaf:

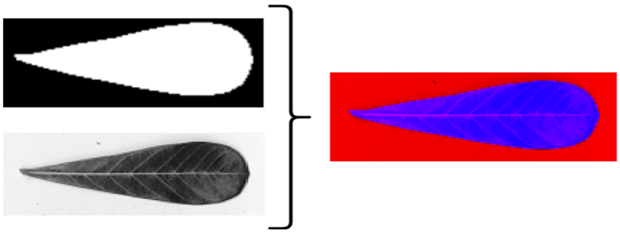

Fig. 2 Binarization of the leaf's image

(ii) Step 2: Extraction of characteristics

An image is a signal. From a mathematical point of view, it is a matrix of numbers representing this signal [16]. It thus contains several semantic information. An important approach to describe a region is to quantify its texture content and a frequently used approach for texture analysis is based on the statistical properties of the intensity histogram. One class of these measures is based on statistical moments. The expression for the nth moment of the mean is given by [17a]:

$$\mu_n = \sum_{i=0}^{L-1} (z_i - m)^n \, p(z_i) \qquad (2)$$

Four texture descriptors based on the intensity histogram were defined as relevant [17b]:
- Mean: A measure of average intensity.

$$m = \sum_{i=0}^{L-1} z_i p(z_i) \qquad (3)$$

- Entropy: A Measure of Randomness.

$$e = -\sum_{i=0}^{L-1} p(z_i) log_2 p(z_i) \qquad (4)$$

- Smoothness: A measure of the relative Smoothness of intensity in a region.

$$R = 1 - 1/(1+\sigma^2) \qquad (5)$$

- Standard deviation: A measure of average contrast.

$$\sigma = \sqrt{\mu_2(z)} = \sqrt{\sigma^2} \qquad (6)$$

Where z_i is a random variable indicating intensity, $p\ (z_i)$ is the function of intensity histogram levels in a region, L is the number of possible intensity levels.

(iii) Step 3: Resampling using the bootstrap technique

The bootstrap technique is a method based on multiple replications. It consists in creating "new samples" from the initial sample using the Wong and Easton algorithm [18]. In our case, the samples per plant are unique because of their originality, but we need a large number of samples per species to make learning more efficient, hence the usefulness of resampling. From forty-one (41) plants, we were able to reconstruct a new database of ten thousand (10,000) plants and extract the values of the selected characteristics.

(iv) Step 4: Resampling using the bootstrap technique

Artificial neural networks (ANNs) are computer systems inspired by biological neural networks. It is based on a set of connected units or nodes called artificial neurons, which loosely model the neurons of a biological brain. Each connection, like synapses in a biological brain, can transmit a signal from one artificial neuron to another [19]. An artificial neural that receives a signal can process it and then signal the other artificial neural connected to it. They automatically generate identifying features from the learning material they process. The initial goal of the ANN approach was to solve problems in the same way a biological brain would.

Before processing the data (neural network application), they were first divided into two sets: a model training set, and a model validation set. The validation set has no effect on training and therefore provides an independent measure of network performance during and after training. It represents 25% of the data and the 75% is attributed to the training set. Subsequently, this training set was divided into three sub-sets: a test set, a validation set and a training set, with a respective percentage of 15%, 15% and 70% [20]. The evaluation of the performance of a neural network model is highly dependent on the distribution of samples in the data set.

Training data are presented to the network during training and the network is adjusted for its error.

Validation data are used to measure the generalization of the network, and to stop training when the generalization stops improving.

Test data has no effect on training and therefore provides an independent measure of network performance during and after training.

The developed prediction model uses a two-layer feedforward network, with a hyperbolic tangent (tanh) transfer function in the hidden layer and a linear transfer function in the output layer. The activation function which combines the functions of the hidden layer and that of the output layer is described by the formula:

$$Output = (LW \times \tanh(IW \times input + B_1)) + B_2 \qquad (7)$$

Where:

- ***Output*** predicts the plant family.
- ***Input*** represents the input parameters.
- ***IW*** is the weight of the connections between the input layer and the hidden layer.
- ***LW*** is the weight of the connections between the hidden layer and the output layer.
- ***B1*** is the bias of the input layer.
- ***B2*** is the bias of the output layer.

With:

$$\boldsymbol{A = LW}$$

$$\boldsymbol{X = tanh\ (IW\ input + B_1)}$$

$$\boldsymbol{B = B_2}$$

The activation function is of the form:

$$Y = AX + B \qquad (8)$$

The ***IW*** weights and the input biases $\boldsymbol{B_{1j}}$ made it possible to calculate the inputs of all the neurons of the hidden layer. Thus, the input of the neuron $\boldsymbol{N_j}$ was calculated by the following formula:

$$\boldsymbol{N_j = \left(\sum_{i=1}^{4} IW_i \times input_i\right) + B_{1j}}$$

$$\boldsymbol{j = 1, \ldots, 10} \qquad (9)$$

- ***j*** is the number of neurons.
- ***i*** represents the input parameter, in our case it is the four texture parameters.

To obtain the value of the neuron output, we applied the activation function. In the case of the output of the neuron N*j*, the activation function is written as follows:

$$\boldsymbol{output} = \left(\sum_{i=1}^{10} \boldsymbol{LW_i} \times \boldsymbol{tanh(Nj)}\right) + \boldsymbol{B_2} \qquad (10)$$

The different values of the neuron output correspond respectively to different types of family. Since our network contains two nodes at the output, our output is of the following form:
(7) ⇔ (10) ***=Y = (**y_1, y_2**)*** = code of plant families (11)

Then we identify the three families of plants according to the following table 1:

Table 1: Coding of plant families for automatic classification

Families' name	Code of plant families
APOCYNACEAE	0 1
RUBIACEAE	1 0
EUPHORBIACEA	1 1

3. Result and Discussion

3.1 Result

We have established a new automatic plant classification technique based on artificial neural networks. This technique has allowed us to develop a model for plant family recognition. This model is based on a mathematical formula deduced from the combination of a hyperbolic tangent function (tanh) and a linear function. We tested this model on a set of 3750 local plants. We evaluate the performance of our model.

The table below lists the plants predicted from those observed.

Table 2: Confusion matrix for local plant base

	Classified plants	Unclassified plants
Positive model test	TP= 2676 **71,4%**	FP= 0 **0,0%**
Negative model test	FN= 0 **0,0%**	TN= 1074 **28,6%**

In the following table, we calculate the basic indicators of the quality of prediction.

Table 3: Performance indicator for the local plant base

Performance Indicator	Taux
RMSE	1,348e-14
R-SQUARE	99,99%
SENSIBILITY	84,85%
SPECIFICITY	100%
PRECISION	100%

3.2 Discussion

The quality of the prediction model is evaluated on the basis of the root mean square error (RMSE) and the coefficient of determination (R-square).

The RMSE is a measure of the differences between the values predicted by the model and the observed values. The RMSE value is 1.348e-14; a very low value indicating that the model is good.

The coefficient of determination indicates the correlation between predicted and observed values. The coefficient R^2 = 99%, a value that tends towards 100%; this shows that the model has indeed ranked the majority of plants, therefore it is an efficient model.

The plant confusion matrix shows the responses of the model after being tested. The number of correct answers (True Positive = 71.4% and True Negative = 28.6%) is high compared to the number of incorrect answers (False Positive = 0% and False Negative = 0%). This shows that the classification of plants by the model corresponds exactly to the different families of plants established by botanists and therefore our model is reliable.

The results of the intrinsic validity of our model are presented below:
Sensitivity: *(TP/ (TP+FN)) = 84.85%*
Specificity: *(TN/(TN+FP)) = 100%.*
Precision: *TP / (TP + FP) = 100%.*
In order to verify the efficiency of our model, we applied it to an international database such as the Flavia database and obtained the following results:
The table below lists the plants predicted from those observed.

Table 4: Confusion matrix for the basis of FLAVIA International Plants

	Classified plants	Unclassified plants
Positive model test	TP= 891 **71,3%**	FP= 0 **0,0%**

Negative model test	FN= 0 **0,0%**	TN= 1074 **28,7%**

In the following table, we calculate the basic indicators of the quality of prediction.

Table 5: Performance values of the model tested on the basis of FLAVIA International Plants

Performance Indicator	**Taux**
RMSE	1,136e-14
R-SQUARE	98 %
SENSIBILITY	84,85%
SPECIFICITY	100%
PRECISION	100%

Finally, to confirm the performances of the model, we compared it to the work of *Aimen Aakif* and Al. who give the best results of the works reviewed.

Table 6: Comparison of accuracy

	ACCURACY	
	LOCAL BASIS	FLAVIA: BASIS
Aimen Aakif	96%	96%
Our Model	**99,99%**	**98%**

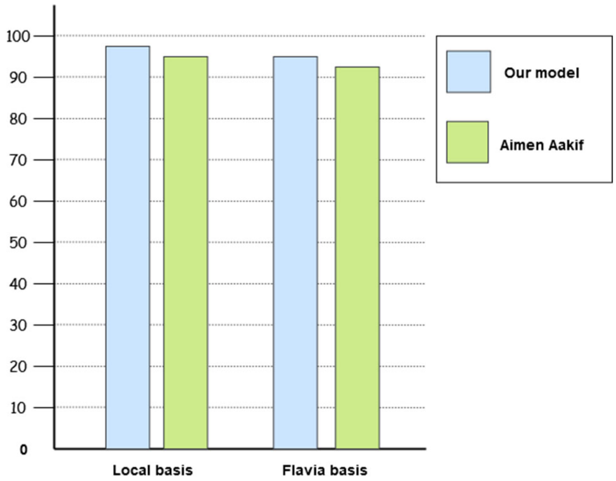

Fig. 3 Comparison of accuracy

4. Conclusion

Automated classification of plant families remains a difficult task in botany. The technique we propose offers a significant discrimination capability to classify plant families with reduced error. This classification of plants by artificial neural networks has allowed us to develop a model for plant family recognition. This model is based on a mathematical formula deduced from the combination of a hyperbolic tangent transfer function (tanh) in the hidden layer and a linear transfer function in the output layer. All the performance indicators namely the error rate (RMSE), the determination coefficient (R^2), the sensitivity, the specificity and the precision of the model show that the leaf texture parameter approach for plant classification offers a good accuracy. The results of this model have been compared with the work of Aimen Aakif who gives the best results of the works reviewed. The accuracy rate of our model is higher than that of our predecessor, Therefore, our model is efficient in predicting the family of a plant. It can guide the botanist in his identification process. Hence, we will make a graphic interface for user, in the future.

References

[1] TREY Zacrada F.O and Al**.** « Classification des espèces végétales par famille » International Journal of Scientific Research & Engineering Technology (IJSET)Vol.7 pp. 1-5 Copyright IPCO-2019ISSN 1737-9296, 2019

[2] TREY Zacrada F.O and Al «Classification of plant species by similarity using automatic learning» https://doi.org/10.1007/978-3-030-41593-8_14 International Conference on e-Infrastructure and eServices for Developing Countries AFRICOMM 2019: e-Infrastructure and e-Services for Developing Countries pp 186-201, 2020.

[3] Alain Coulon, Artificial Intelligence Supplement, Letter 111 ; Spring

[4] Vijay satti and Al, an automatic leaf recognition system for plant identification using machine vision technology, International Journal of Engineering Science and Technology (IJEST), 2016.

[5] Irié A. Zoro BI and Al. « Caractérisation botanique et agronomique de trois espèces de cucurbites consommées en sauce en Afrique de l'Ouest : *Citrullus* sp., *Cucumeropsis mannii* Naudin et *Lagenaria siceraria* (Molina) Standl», *BASE* [En ligne], Volume 7 (2003), Numéro 3-4, 189-199 URL : https://popups.uliege.be/1780-4507/index.php?id=14375, 2003

[6] Munisami and Al, «Plant Leaf Recognition Using Shape Features and Colour Histogram with K-nearest Neighbour Classifiers», Article in Procedia Computer Science, DOI: 10.1016/j.procs.2015.08.095, December 2015.

[7] Arun priya and Al, an efficient leaf recognition algorithm for plant classification using support vector machine**,** Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, March 21-23, 2012.

[8] Dimitris G. Tsolakidis and Al., « Plant Leaf Recognition Using Zernike Moments and Histogram of Oriented

Gradients », SETN 2014, Springer International Publishing Switzerland 2014, LNCS 8445, pp. 406–417, 2014.

[9] Shubham Lavania, Palash Sushil Matey, Leaf Recognition using Contour based Edge, Detection and SIFT Algorithm, IEEE2002

[10] Alexis Joly and Al, LifeCLEF 2016: Multimedia Life Species Identification Challenges, CLEF: Cross-Language Evaluation Forum, Évora, Portugal. pp.286-310, ff10.1007/978-3-319-44564-9_26ff. ffhal-01373781f, Sep 2016

[11] Abdul Kadir and Al, Leaf Classification Using Shape, Color, and Texture Features, International Journal of Computer Trends and Technology- July to Aug Issue 2011.

[12] Jagadeesh and Al, Image Processing Based Detection of Fungal Diseases in Plants, Procedia Computer Science 46:1802-1808, DOI: 10.1016/j.procs.2015.02.137, December 2015.

[13] Youcef Djeriri, Les Réseaux de Neurones Artificiels, UDL-SBA-2017, 20 September 2017

[14] **[a, b]** Aimen Aakif and Muhammad Faisal Khan, Automatic classification of plants based on their leaves, ScienceDirect, 2015.

[15] Nobuyuki Otsu and Al, A Threshold Selection Method from Gray-Level Histograms, IEEE Transactions on Systems, Man, and Cybernetics (Volume: 9, Issue: 1, Jan. 1979) DOI: 10.1109/TSMC.1979.4310076, Page(s): 62 – 66

[16] Elise Arnaud - Edmond Boyer, Analyse d'images - introduction, http://perception.inrialpes.fr/people/Boyer/Teaching/L3/

[17] **[a, b]** Rafael C. Gonzalez, Digital image using MATLAB, chapter 11 p 466.

[18] Wong, C. K. and M. C. Easton. An Efficient Method for Weighted Sampling Without Replacement. SIAM Journal of Computing 9(1), pp.111–113,198

[19] www.f futura-sciences.com ' Planet ' Definitions Alain Coulon, Artificial, Intelligence Supplement, Letter 111; Spring 2018.

[20] Sakshi Kohli, Surbhi Miglani, Rahul Rapariya. basics of artificial neural network, IJCSMC 2014, 3, 745-751.