

PAPER • OPEN ACCESS

Recognition of Arabic Handwritten Diacritics using the new database DBAHD

To cite this article: N Lamghari and S Raghay 2021 *J. Phys.: Conf. Ser.* **1743** 012023

View the [article online](#) for updates and enhancements.



The Electrochemical Society
Advancing solid state & electrochemical science & technology

240th ECS Meeting ORLANDO, FL

Orange County Convention Center Oct 10-14, 2021



Abstract submission due: April 9

SUBMIT NOW

Recognition of Arabic Handwritten Diacritics using the new database DBAHD

N Lamghari¹ and S Raghay²

¹ ENSA, LIPIM, Sultan My Slimane University Khouribga, Morocco

² FSTG, LAMAI Cadi Ayyad University Marrakesh, Morocco

Email : n.lamghari@usms.ma

Abstract. Recognition of handwritten Arabic characters is gaining momentum and research in this area has increased considerably in recent years. However, research remains modest compared to that performed in other scripts. This is mainly due to the morphology of Arabic writing, in particular its richness in diacritical marks. These signs are generally recognized by adopting structural or morphological measures. However, the difficulty and variability of handwriting can sometimes be misleading, thus influencing the results obtained. This article presents a new database for Arabic handwritten diacritics (DBAHD). It is designed to serve the Arabic handwriting recognition systems based on segmentation and machine learning.

1. Introduction

Today, we still have a large number of handwritten paper documents. We often need to represent these documents in digital form in order to be able to use them effectively without having to enter them manually. So there is a large number of old documents kept in the archives. It is therefore important to preserve this endangered heritage, to make it accessible to everyone and to interpret it easily.

Digitization is the solution, but it only provides images of documents, which is not always enough. Indeed, it is often necessary to access the contents of the scanned documents and modify them if necessary.

This is the purpose of optical document recognition systems. So handwriting recognition is an active area of research in pattern recognition and has many practical applications. This is an extremely difficult task due to the wide variations in handwriting styles. In some limited areas, the task of handwriting recognition may become less demanding. For example, in postal code recognition, the problem can be reduced to recognizing a sequence of isolated numerical digits. Even in this limited field, however, many difficult questions arise. One of them is the classification of diacritical marks which constitute an additional challenge for a character recognition system. These are signs that we



associate with the main bodies of a few characters. Also, Arabic is one of the languages that use these signs. This characteristic is one of the main causes of the delay in the recognition of Arabic handwriting compared to other scripts. In fact, according to the bibliographic study that we have done, there is no database dedicated to these signs.

Thus, this article aims to present a new database of Arabic Handwritten diacritics (DBAHD), designed to cover five forms of Arabic diacritics: a point, two points, three points, hamza and medda. In the following section, we present, analyze and study Arabic writing and especially the diacritics. Section III describes related work in the field. Our new database of Arabic manuscripts is covered in sections 4 and 5.

2. Arabic script and diacritics

The most common theory admits that Arabic script is descended from Nabatean script, the direct ascendant of ancient Aramaic script which is an offshoot of the Phoenician alphabet. The Nabataean alphabet from which the Arabic alphabet is derived was not provided with diacritical marks [1][2]. Arab philologists have created secondary components to differentiate letters that have the same shape but different phonetics. These components are called diacritical marks and can be points (one, two or three), vocalization signs or other signs such as hamza, chadda or madda, wasla, etc. So the number of letters in the current alphabet is 28 letters (figure 1).

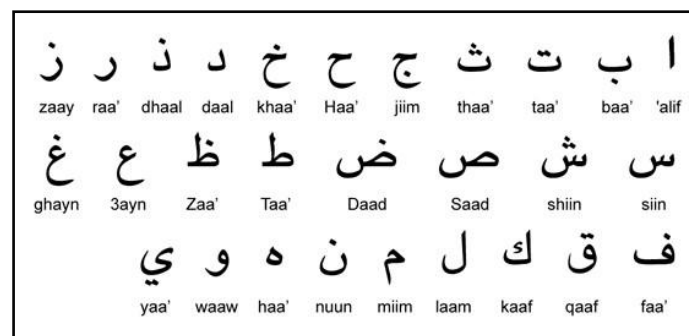


Figure 1. Arabic alphabet

2.1. Diacritical point

Diacritical points appear above or below the character only. The maximum number of points that may have a letter is three points above the character or two points below. There are 15 letters of the Arabic alphabet with dots. Figure 2 illustrates the use of diacritics to differentiate letters.

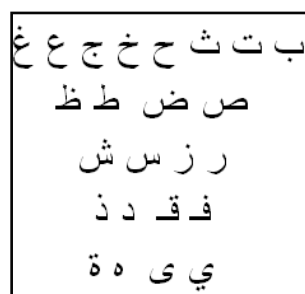


Figure 2. Letters with the same body and a number and / or position of different diacritical points

2.2. Signs of vocalization

In Arabic, vowels are not letters but diacritical marks associated with the letters to which they apply [3]. There are three types of vowels: long vowels, short vowels and doubles. Long vowels are the three letters: "ا، و، ي" which are both vowels and consonants. As vowels, they allow prolonged pronunciation of letters. On the other hand, when they do not play the role of vowels, they behave like the rest of the consonants, and can therefore include diacritical marks. The short vowels are the four signs: "fatha", "damma", "soukoun" and "kasra". The signs "fatha" (small bar) and "damma" (small و) are placed above the letter and are pronounced respectively as an "a" and an "o" in French. The "soukoun" sign (small o), which is a small circle above the letter indicates that no sound exists in addition to the consonant. As for the "kasra" sign (small bar), it is placed below the letter and is pronounced like an "i" in French. Double vowels are, as their name suggests, double signs (short vowels). They are placed on the last letter of the word. There are three double vowels: double "fatha", double "kasra" and double "damma". Generally, short and double vowels are not written (except in textbooks). They are implicit and it is up to the reader to guess them.

As specified by Y. Bahou et al in [4], vowels are very important in Arabic and their absence can be confusing, since the word can change its meaning if the correct vowel is not in its place. The figure below (figure 3) shows examples.

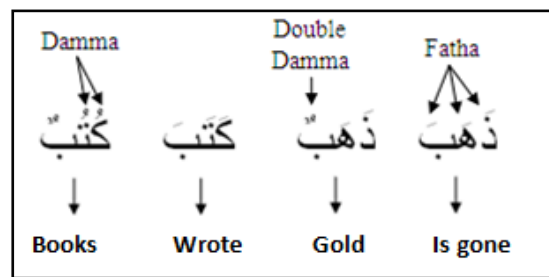


Figure 3. Importance of vowels in Arabic grammar

2.3. Other diacritical signs

The other diacritical signs are: hamza, chedda, madda, wasla. The chedda is an accentuation of the letter: instead of writing a letter twice, we place this sign above the letter. The hamza sign is both a consonant and a diacritical sign. When not placed on a letter it acts as a consonant and may contain vocalization signs. Otherwise, it is placed above the following letters: "ا، و، ي" or below the letter "ا". The sign madda (prolongation) is placed on the letter "ا" to indicate that this letter takes the place of two consecutive "ا" or that this letter must not bear the hamza. The wasla sign is placed on the first letter "ا" of words whose determinants are defined (preceded by "ال") and which are in the middle of a sentence. Figure 4 below shows an example of words containing these signs.

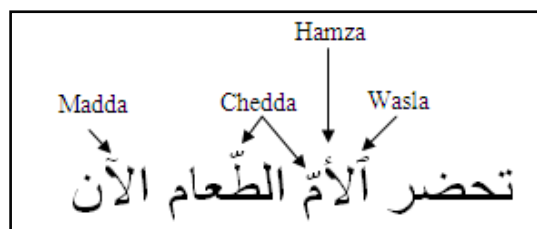


Figure 4. Other diacritical signs

3. Related work

The literature review we have made in this area, we unveiled the lack of work on the study of the Arabic diacritical signs. Indeed, most of these works deal with these signs combined with the Arabic characters which are assigned to it. The few works that try to recognize diacritical marks as

independent characters proceed according to structural methods. They are based on mathematical and geometric measures. Nevertheless, the Arabic handwritten diacritical marks do not differ from the Arabic handwritten characters. Also, they are complex, have different forms and depend on the writers. In this section, we describe some work in the field.

In [5], the researchers first extract the pseudo words and then segment them into characters. Each character "i" is analyzed to qualify it as a basic component or a diacritical mark. Thus, this analysis is based on the measurement of the size of the component and the distance between the centroid of the component and the estimated baseline. In the second step, the main components selected in the first step are refined. The idea is to find the components above or below other larger components. This is done by calculating the overlap of the bounding boxes for the blobs.

The secondary components (diacritical marks) of a PAW can be, according to the authors, points, Shaddah, Hamza, part of a broken character, etc. Consequently, any secondary components possibly written in points must be identified. The morphological properties of the components are used to identify the points. They estimate the location of each secondary component from the baseline of the PAW image. Thus, a secondary component is above the main body of the PAW, if its center of gravity is above the baseline. It is below the main body if its centroid is below the baseline. For the classification of the secondary components, the authors base themselves on the analysis of the width and the height of the bounding box. Thus, the Shaddah signs and three points grouped together are recognized using a polygonal approximation. Researchers use several heuristic methods to extract diacritical marks and signs.

And in another study [6], researchers rely on the analysis of diacritical marks for the identification of writers. They first apply a preprocessing step, then extract all the diacritical marks and calculate the histogram of the local binary pattern for each diacritical mark. Each diacritic has a histogram of length equal to 256, which will give a vector of length equal to $256 \times n$, where n is the number of diacritics. For classification, they use the nearest neighbor K method. They first calculate the minimum diacritical distance for the same author and then add these distances. Diacritical marks were extracted from the IFN / ENIT database. 120 diacritics were used for learning and 80 for testing. These were divided into 16 parts, starting with 5 diacritics and then gradually adding 5 diacritics. They run the experiment for a set of 287 editors and get an identification rate of 51.22%. They conclude that this low rate is due to the fact that LBP is invariant to monotonic changes in the grayscale image as in binary images. By excluding this value from the histogram they get an average recognition rate of 97.56%.

In [7], the researchers treat four types of diacritics, namely: '·', '..', '::', 'ء'. They recognize these secondary components using the following parameters: the number of segments in the horizontal direction (height) and the number of elements and columns in each segment, as well as by the connectivity analysis. The number of points is estimated using two methods. In the first, we estimate the number of extreme contours by removing the outline of the primary part. As for the second method, it is based on estimating the total area of the points of a character and then dividing it by the average area of a point. The latter is pre-calculated based on the areas of the characters used in the learning phase. They finish by focusing on the presence or absence of a hole, the number of holes and the elongation of a hole which can provide additional information to resolve certain ambiguities. They do not present any concrete results in relation to the recognition of diacritical marks.

However, as detailed in this literature review, no database of diacritics is proposed in the field, given that most use structural and morphological methods to recognize these signs. Indeed, to develop a recognition system based on segmentation and learning techniques, a database covering Arabic handwritten diacritical marks is necessary to facilitate the recognition process. Given this need, we offer in the following section our database of Arabic manuscript diacritics.

4. Our database of Arabic handwritten diacritics

To develop a system of recognition of Arabic handwriting characters based on segmentation, a database of diacritical marks is necessary to facilitate the recognition process. This database must cover all forms of handwritten diacritical marks. Unlike other scripts whose only dots are diacritics, Arabic uses other diacritics as shown above. It is in this context that we deemed it necessary to

develop a database of diacritical marks which will be used in the field of recognition of Arabic handwriting.

4.1. Alphabet diacritics used

We limit our study to 3 types of diacritical signs: points, hamza and madda. The madda sign is always placed at the top of the letter „ا“. As for the dots and the hamza sign, they can appear above or below the letters. The hamza sign is placed above the letters “ا, ي, و” and below the letter “ا” the maximum number of points you can have is three points above the letters “ث, ش” and two points in below the letter “ى”. When writing, scripters can sometimes combine two or three dots into one form. Figure 5 shows examples of different diacritical points present in the IFN / ENIT database [8]. For this, we define an alphabet of diacritics defined as follows:

- P: the case of a single point, two ungrouped points, three ungrouped points
- H: represents the hamza sign
- M: represents the sign madda
- T: two points grouped together
- V: three points grouped together

Number of points	Examples							
1								
2								
3								

Figure 5. The different forms of the diacritics points

4.2. Acquisition and storage

The data in this database, in particular the samples concerning the following signs: hamza, one point, two points and three points are extracted from the IFN / ENIT database [8]. For the madda diacritical mark, we preferred to collect samples of this sign ourselves, since the IFN / ENIT database does not contain enough examples containing this sign. Thus, 50 writers wrote the letter ‘ا’ twice with the sign Medda.

As the data stored in the IFN / ENIT database is already binarized and pre-processed, we only process diacritics of the "Medda" type. The forms were scanned with a quality of 300 dpi. And to reduce processing time and to do as much automatically as possible, a program has been developed with Matlab to automatically extract medda signs from different forms. After collecting and storing the images, conventional preprocessing tasks were performed such as: noise filtering and binarization of the images. To adjust and unify the sign size, the sign images have been resized and normalized to 80x80. Due to the writing errors, a manual verification is required. For this, we verified the medda signs extracted from the collected forms. Then, we saved the images of the same class in a separate folder with its own name.

Thus, the DBAHD database consists of 500 diacritics (one point, two points, three points, hamza, madda). It contains five folders; each folder corresponds to a type of diacritical sign and contains 100 examples of this sign. Figure 6 below provides an overview of this database.

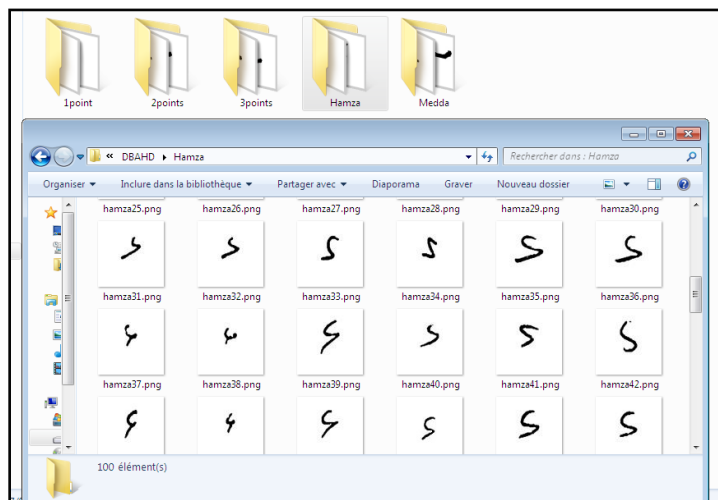


Figure 6. Overview of Database for Arabic Handwritten Diacritical DBAHD

5. Working with our database

The DBAHD database is publicly available for the purpose of research. It is available at: <https://github.com/nidaliso/DBAHD/>. The database consists of five folders: "1point", "2points", "3points", "Hamza" and "Medda" which respectively correspond to diacritics: "P", "T", "V", "H" and "M" above mentioned.

With this database, it is possible to develop and test systems for recognizing handwritten characters in Arabic. This database complements the new DBAHCL database that we presented in the article [9]. It is a database of Arabic handwritten characters without diacritics. To report a benchmark recognition performance on the developed database, we used two methods: structural and connectionist. In the structural approach, we implement an algorithm based on the morphological characteristics of said diacritics. And in the connectionist approach, we use a feed forward neural network.

5.1. Identifying diacritics using a structural approach

The algorithm that we define below is used in addition to another method of detection of diacritical signs that we applied before in [template]. So we used a modified version of the method of Parvez et al. [5]. This method consists in extracting the connected components based on the area. Thus, the component having the maximum area is identified as the main body of the character (MBC). The information is extracted in relation to this main body. All other components of the image are called diacritics. The location of each diacritic can be estimated from the centroid of the main body of the character. A Di diacritic is above the main body of a MBC character, if the Di centroid is located above the MBC centroid. Di is below the MBC if the centroid of Di is below the centroid of the MBC. We recognize diacritics based on 3 tests. The first test checks whether the sign is a zigzag (hamza). The second test consists of recognizing the three points grouped together in the form of a 'V' 'upside down, by studying the concavity of the outline of the diacritical mark. The third test can detect signs extended horizontally whose length is greater than a threshold 's'. These signs can be either "M" (medda) or "T" (two dots grouped together). The sign is recognized "M" if it is above the MBC and the height H of MBC is much larger than its width L, in other words, it is a question of checking whether the MBC corresponds to 'l'. Since the sign "M" can only be on the character alif 'l', recognition of this sign then amounts to recognizing the body of the character associated with this sign. The sign is recognized as "P" (period) if it does not meet any of the three tests. Figure 7 below summarizes these three tests.

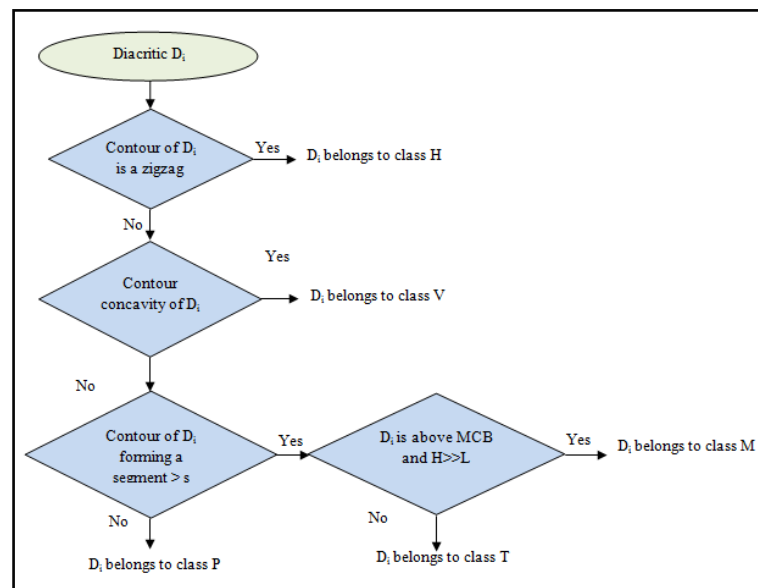


Figure 7. Diagram of our identification diacritics algorithm

Experimentation with the diacritics identification algorithm described above has revealed some problems, particularly for the three points. In a few cases illustrated in Figure 8 below, we have obtained a total of points exceeding three. To remedy this problem, we set the maximum points to remember at three. After this modification, testing our program on the DBAHD database made it possible to obtain a recognition rate of 100% of the signs Hamza and Madda, 98% for 3 points, 92% for the two points and 93% for a single point.

We adopted this algorithm to extract the characteristics relating to the diacritics. Also, we have obtained very encouraging recognition rates which vary from 95.89% [10] and 98.27% [11].

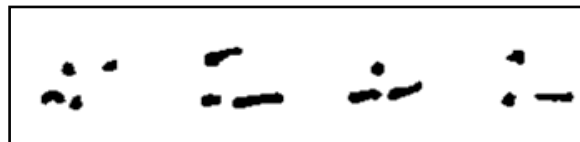


Figure 8. Example of a case where the result is greater than three points

5.2. Identification of diacritics with a connectionist method

The neural network is a static classifier that requires vector functions of fixed size. For this reason, it has been successfully used in the classification phase of handwriting recognition systems. However, compared to other approaches, it requires more calculations [12].

In this context, we create a neural network with back propagation using the MATLAB neural toolbox. The data is divided into three parts: 70% of the isolated characters from our DBAHD database are used for training ($0.7 * 500$), 15% for the test phase and 15% for validation.

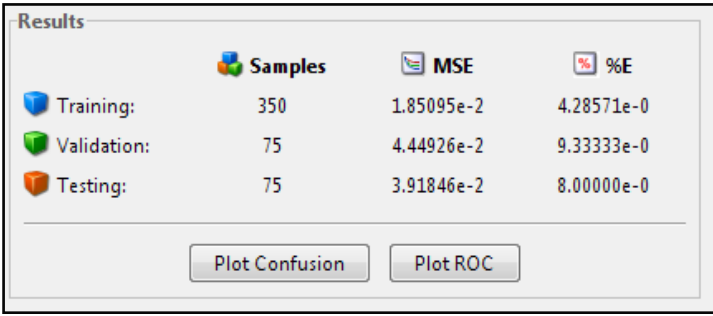
The network used has an input layer, an output layer and a hidden layer. The input layer receives the vector resulting from the extraction phase of the characteristics of the diacritic to be recognized. The hidden layer is an intermediate layer formed by hidden neurons. These neurons calculate the distance between the input vector and all of the training vectors. And the output layer provides the result of the recognition. To function properly, all the data of the neural network must be standardized. Indeed, standardization is essential for multilayer neural networks because it makes it possible to maintain network weights at relatively small intervals, to optimize learning conditions and at the same time to improve convergence [13].

The input layer of our neural network includes 100 neurons which correspond to the set of characteristics extracted from: pixel density, resizing, invariant moments and regional characteristics.

The output layer is composed of 5 neurons which correspond to the number of classes. The output vector is defined with a variable called "target". All target vectors are 5 elements which contain a "1" at the position (line) of the diacritic and a "0" everywhere else. For example, the ' ' which is the first diacritic of our target is represented as a vector containing 1 in the first element and 0 in the 4 other elements.

For the hidden layer, there is no hard and fast rule for determining exactly how many neurons to use in this layer [14]. However, some commonly accepted heuristics advance different numbers, concerning the number of hidden neurons [13]. But, to get reliable results, it would be better to try different sizes. Indeed, the number of hidden nodes varies according to the applications and this number must be determined experimentally [13]. After several attempts to adjust the number of hidden layers, starting with 10 neurons, we found that from 20 hidden neurons, the process has converged. Indeed, a network of 23 neurons in a hidden layer could better predict the characters with a very small error. Thus, after having trained the network four times ($4 * 1000$ epochs), we obtained a very encouraging small error: $2.55e-2$ with an average recognition rate of 94.4% (Figure 9).

Of course, the rate obtained using the structural method is more interesting. However, recognition of the "Medda" diacritic implies prior recognition of the "Alif" character to which it is assigned. This is not always possible. While with the connectionist method, we do not need to recognize the main body of the character.



	Samples	MSE	%E
Training:	350	$1.85095e-2$	$4.28571e-0$
Validation:	75	$4.44926e-2$	$9.33333e-0$
Testing:	75	$3.91846e-2$	$8.00000e-0$

Figure 9. The results obtained from Our Neural Network

6. Conclusion

The field of recognition of Arabic script is still lacking in the absence of a reference database covering all forms of diacritical marks. Such a database makes it possible to compare the results of the different systems for recognizing handwritten Arabic characters. It is in this context that we have introduced, in this paper, a new database which covers five forms of Arabic diacritical marks. **This database contains 500 forms of diacritics and can be used to compare the effectiveness of certain techniques and approaches for recognizing handwritten Arabic characters which contain diacritical marks, in particular those which recognize the word after segmentation. In perspective, we plan to expand this database to include other signs such as: 'chedda' and 'wasla' and also vocalization signs.**

7. References

- [1] Djebbar A 2012 *History and evolution of the Arabic language*, http://www.lescahiersdelislam.fr/Histoire-et-evolution-de-la-langue-arabe_a137.html
- [2] Zghibi R 2002 Le codage informatique de l'écriture arabe : d'ASMO 449 à Unicode et ISO/CEI 10646 Digital document 2002/3 (Vol. 6) pp 155-182. DOI 10.3166/dn.6.3-4.155-182
- [3] Menasri F 2008 *Contributions to the recognition of Arabic handwriting* (Paris).
- [4] Bahou Y, Aloulou C, Hadrich Belguith L and Ben Hamadou A 2006 Adaptation and implementation of hpsg grammars for the analysis of unvoiced Arabic texts *Proc. of the 15th Pattern Recognition Congress and Artificial Intelligence (RFIA'2006)*.
- [5] Parvez M. T and Mahmoud S. A 2013 Arabic handwriting recognition using structural and syntactic pattern attributes *Pattern Recognition*, 46(1) pp 141-154.

- [6] Lutf M, You X and Li H 2010 Offline Arabic handwriting identification using language diacritics *Proc. Int. Conf. on Pattern Recognition* pp 1912-1915.
- [7] Ouchtati S, Ramdani M and Bedda M Un réseau de neurones multicouches pour la reconnaissance hors ligne des caractères manuscrits arabes *Sciences & Technologie. A, sciences exactes*, no 17 pp 99-105.
- [8] Pechwitz M, Snoussi Maddouri V, Märgner V, Ellouze N and Amiri H 2002 IFN/ENIT Database of Handwritten Arabic Words *CIFED'02* pp 129-136.
- [9] Lamghari N, Raghay S 2017 DBAHCL: Database for Arabic Handwritten Characters and Ligatures *International Journal of Multimed Information Retrieval* pp 1-7, Doi: 10.1007/s13735-017-0127-x
- [10] Lamghari N, Charaf M E H, Raghay S 2016 Template Matching for recognition of handwritten Arabic characters using structural characteristics and Freeman code *The International Journal of Computer Science and Information Security* 14(12) pp 31-40.
- [11] Lamghari N, Charaf M E H, Raghay S 2017 Hybrid Feature Vector for the Recognition of Arabic Handwritten Characters Using Feed-Forward Neural Network *Arabian Journal for Science and Engineering* pp 1-9. Doi:10.1007/s13369-017-2969-1.
- [12] Mars A and Antoniadis G 2016 Arabic online handwriting recognition using neural network *International Journal of Artificial Intelligence and Applications (IJAIA)* Vol. 7 No. 5.
- [13] Gnana Sheela K and Deepa S. N 2013 Review on Methods to Fix Number of Hidden Neurons in Neural Networks *Hindawi Publishing Corporation Mathematical Problems in Engineering* doi: <http://dx.doi.org/10.1155/2013/425740>.
- [14] Zhang G P 2000 Neural Networks for Classification: A Survey *IEEE Trans on Systems, Man and Cybernetics Part C*, vol.30, no. 4, pp 641-662.

Acknowledgments

We acknowledge the support provided by the members of the process engineering, computer science and mathematics laboratory of the Khouribga National School of Applied Sciences, Sultan My Slimane University, Morocco and the members of the laboratory of applied mathematics and computer Science, Faculty of Science and Techniques, Cadi Ayyad University, Marrakesh, Morocco.