Design Engineering

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

An OCR for Arabic Character Recognition with Ensemble Approach based Feature Selection for Enhanced KNN Classification

Ashiq V M
Research scholar,Dept of Computer Science,
Karpagam Academy of
Higher Education,Coimbatore
vmashiq@gmail.com

Dr E J Thomson Fredrik
Professor,Dept of Computer Science, Karpagam
Academy of Higher Education,Coimbatore
Thomson500@gmail.com

Abstract:

Data input issues, which seem to be a barrier for the data computing sector, could be solved with an Optical-Character-Recognition (OCR) mechanism. As a result, OCR mechanisms have already been designed virtually for all languages in the world, including Arabic. Over the last 30 years, much fundamental research has gone into the creation of an effective Arabic-OCR (AOCR) mechanism. Because of the increasing quantity of materials accessible on the website, in email accounts, and online databases, documentation categorization has become a necessary job. It's usually accomplished following proper selection of features, which entails choosing suitable characteristics to improve the accuracy of classification. The large percentage of feature-based textual classifying techniques depend on creating a term-frequency and inverse-document-frequency based on features representation, which is inefficient across many cases. Furthermore, many content categorization research is concentrated on the English-language. Despite the difficulty of the Arabic-language, this research paper concentrates on AOCR, which has received less attention. The "Extended Particle Swarm Optimization" (EPSO) methodology is introduced for selecting optimal features from extracted features to comply with Arabic-Character-Classification, and the "Enhanced K-Nearest Neighbor" (EKNN) has been employed as a classification model for identifying and classifying or simply recognizing the particular Arabic character in this research article. Various assessment metrics, comprising Accuracy, Precision, and Recall, are utilized to evaluate this method. Experimentation on an actual dataset is also carried out, and also a comparative with existing Deep-Belief-Networks (DBN) techniques had performed.

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

In comparison to the DBN technique, the developed EPSO-EKNN mechanism obtains a higher accuracy value. The findings show that the recommended selection of features by EPSO is effective in increasing AOCR accuracy.

Keywords: Optical Character Recognition, Particle Swarm Optimization, Enhanced-KNN, Arabic Text Classification

I. INTRODUCTION

An OCR comprises computer interaction with the actual language of the human. The OCR's primary goal aims to allow machines to create substantial natural or human language content to render the text-composing data accessible to many machine learning and data mining techniques. Some of the best data mining and machine learning applications are Text-Recognition (TR) and Character-Recognition (CR). This involves classifying simple text materials into specified groups [1].

Numerous scientists have been drawn to this CR because of the huge collection of publications found on the website, in a mailbox, and in academic resources, as well as its primary usage in several fundamental practical systems such as narratives classified by subjects, scholarly articles, and online sources in library collections, which are typically organized by professional areas and sub directories including such spam detection, which categorizes email. Automation of CR is essential for reducing the cost and time while also reaching higher recognition accuracy. To continue improving the performance of the classifier, both the CR and conventional classifying problems require data preparation [2].

Several approaches for various languages, such as Chinese, English, and Japanese, have already been built-in CR mechanisms. The Arabic-language, on the other hand, had also shown only very slight development. As a result, Arabic-CR remains a prevailing and largely unsolved study area. The digital revolution of Arabic docs could even make way to cultural and Islamic text handling (information retrieval, exploring, etc.) [3].

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

Row	challenges	Description			
1	Different shapes	ع عـعـ ـع	هـهـ ـه	ك ـك	
2	Secondaries exists or not	ظ ـظ	غ ـغ	ق ـق	
		ش ـش	ض ـض	ز ـز	ذ ـذ
3	Number and position of secondaries	ح ج خ		ي ت ب	
4	Secondary types	ـشـ ـيـ	ي ـي	ـسـ ـشـ	أ ـك
			ت ـت		

Figure 1: Variation in Arabic Characters

Attempts to digitalize Arabic texts in the past have run into many problems. To begin, the alphabets consisted of 28 letters, each with a different kind and amount of dots, such as 1, 2, or 3 dots. Also, every character has a different style of writing depending on where it appears in a text (start, center, or final) [4]. As a result, every character contains approximately 80 different writing forms or patterns. Over the 1st row, Figure 1 illustrates various forms for the very same Arabic-character. Every combination of characters is shown in the 2nd row with a distinct form of a dot. The figure's 3rd row featuring the characters most with various dot positions. The final row depicts various forms for the identical character depending on its dot form.

The dimensionality minimization of retrieved features is used in the vast of the methods proposed for dealing with Arabic characters. The originating features should always be converted to some other domain in certain techniques (like Principal Component Analysis). Feature-selection techniques are employed to overcome this restriction, even though such approaches are not dependable for limiting back the more useful features. Numerous swarming approaches were being developed to enhance the process of finding the much more relevant features in feature-selection algorithms [5].

Before the classification stage, Feature Selection (FS) is an important stage. It entails identifying a subset of important (major) features to be used in the classifying stage. Due to the obvious massive amount of the data, it is highly recommended preceding TR [6]. Its major benefits

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

comprise making data easier to comprehend, decreasing time for training, and avoiding the dimensions constraint. Furthermore, the complexity of the classification and computational needs (such as storage) will also be decreased [7].

Yet, research on Arabic-CR, particularly those using FS methods, has been only a few. Is therefore related to the complexity of the structural elements in Arabic. In Arabic-texts, there have been 2 categories of features: exterior and interior [8]. Exterior features have been described as aspects that are unrelated to the document's content, such as the author's title, date of publication, and so on. Interior features, on the other hand, are textual lingual features such as lexicons and grammar character traits. Whenever dealing with AOCR, the FS seems to be a multidimensional issue that necessitates an effective optimizing method [9].

The research's problem statement is really to tackle well with the difficulty of character identification in the Arabic-language. Until now, Arabic has been a widely used language. Over one billion people are using the Arabic text around the universe, according to estimates. If OCR mechanisms for Arabic letters become accessible, people will then have significant economic benefits. Considering its cursive form of Arabic writing, therefore, the creation of Arabic OCR mechanisms poses several technological challenges, particularly during the FS and categorization stages. Despite the efforts of numerous academics to find clarifications to the issues, only limited improvement has yet been achieved. A few analysts concentrated on feature extraction, smoothing, and feature matching, whereas others concentrated on conditioning the OCR mechanism with learning algorithms.

The research is motivated by OCR, which has grabbed academics' attention not alone account of the difficult aspect of the issue of closing the readability gap between computers and people, but additionally seeing as it enhances human-machine communication in a wide range of applications. Business context, cheques validation, and a wide range of financial services, industry, and clerical work applications are just a few examples. Since Normal English letters were segregated from each other as well with spacing and wouldn't have the additional complications inherent with Arabic characters, many widely viable OCR solutions are mostly for typed English letters. Maybe that's why, as compared to Arabic, English OCR methods and processes are simpler and more sophisticated. That's also attributed to the reason as numerous characters possess

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

equivalent forms, with variances arising from changes in the location of dots compared to the character's primary component.

The contribution of the research is mostly concentrated on FS and Classification. The EPSO technique is employed as the FS in this case to pick the best AOCR features. For the English language as almost an FS, many methods have been employed. Within our comprehension, this may be the sole analysis of its kind on an Arabic dataset that uses the "EPSO" method for FS. Then, to classify the data, an enhanced version of KNN "EKNN" was employed. Although the KNN is a simplistic and quick classifier, its method and architecture may be improved. The KNN is adjusted with weight characteristics in the enhanced approach. Depending on the types of training sets, the k's factors must be modified. The same frequency of classification would be addressed by adjusting the k's value.

The remaining sections of this research article are organized by following sections: Section 2 discuss some recent articles related to the problem of OCR, Section 3 details about the proposed methodologies module by module also with crisp details of an existing method, Section 4 shows the results and comparison obtained for both existing and proposed methods with different parameters and finally Section concludes this research article.

II. RELATED WORKS

To enhance the precision of clusters in Web pages, the researchers of [10] devised a Hybrid-FS technique. CHI^2, term-frequency, inverse-document frequency, and mutual-information make up this integrated FS technique. In the famous Arabic online magazines, the researchers employed the K-means clustering method. The quality has increased by 28%, according to the researchers.

In [11], Azuraliza, A.B., Siti Rohaidah, A., Nurhafizah Moziyana, M.Y., Yaakub, M.R.,(2017) stated that the researchers presented a novel FS technique for Sentiment-Analysis based on Ant-Colony-Optimization (ACO). Leveraging customer feedback datasets, researchers evaluated the effectiveness of this suggested approach employing a KNN classification. Information-Gain (IG), Genetic-Algorithm (GA), and Rough-Set-Attribute-Reduction (RSAR)

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

were used to contrast the findings. The researchers suggested the highest findings, with improved accuracy of 0.914.

in [12] Mudhsh MA, Almodfer R (2017) proposed a technique for the identification of Handwritten Arabic numbers and letters, the researchers presented an Alphanumeric very Deep-Neural-Network. 13 convolution-layers, 2 max pooling-layers, and 3 completely connected-layers were used to build a classification model. To reduce the number of parameters, two normalization techniques have been used: Augmentation and Dropout. Testing has been performed on two datasets: the AD-Base dataset (a dataset of Handwritten Arabic values from 0-9) and the HACDB dataset (a dataset of Characters with Arabic Handwritten). This approach has a 99.67 percent accuracy for the ADBase dataset and a 97.42 percent accuracy for the HACDB dataset.

In[13] Younis K (2018) created a CNN for recognizing the Arabic handwritten characters. 3 convolutional-layers were suggested, accompanied by completely connected-layers in their suggested CNN. By using AHCD and AIA9K databases, testing findings revealed that the CNN was capable of reaching 94.8 percent and 94.9 percent accuracy, correspondingly.

To enhance the findings, in [14] Peng, C., Limc, S., Chin Neoh, L., Zhang, K., Mistry, K., (2018) created a hybrid based FS by merging this with Simulated-Annealing. They employed 11-regressions and 29-classifications databases to verify the new approach and contrasted it to current techniques. These outcomes are positive.

III. METHODOLOGIES

3.1 Existing model (dbn)

After appropriate pre-processing, a feature learning algorithm is needed to extract a set of sparse features which can be used later to efficiently code the images on feature-space. To achieve that, in [15] Hasasneh, Ahmad & Salman, Nael & Eleyan, Derar. (2019) a model was developed with a novel unsupervised machine learning method termed Deep Belief Networks (DBNs) [15]. These generative methods are constructed of multiple Restricted Boltzmann Machines (RBMs) layers of binary and Gaussian units. These units correspond to the hidden layer or features detectors. Since data is already normalized, the input layer to the first RBM corresponds to zero-

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

mean Gaussian activation values and is often used to compute the first hidden layer as shown in Figure 2 (left).

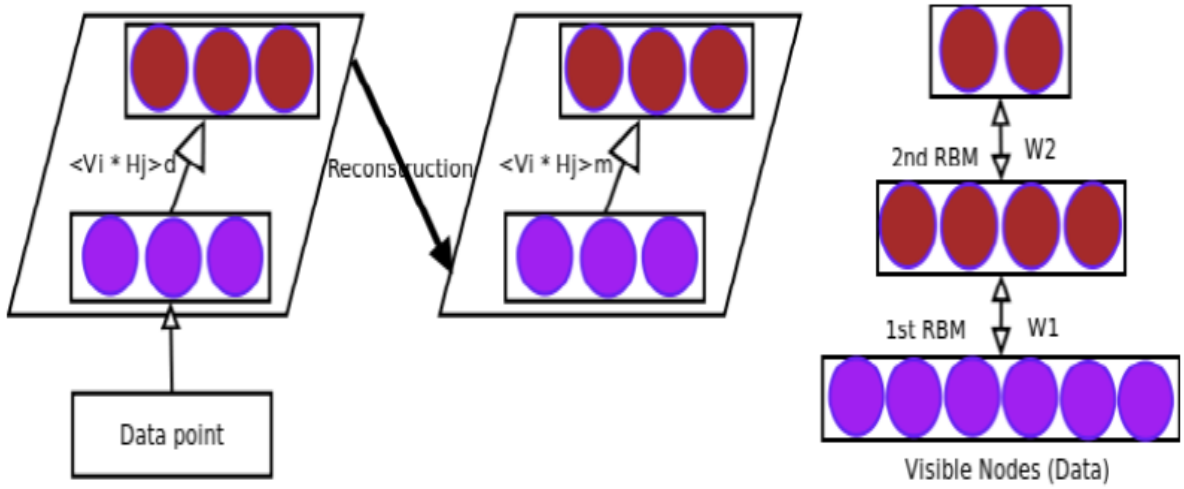

Figure 2: DBN Architecture

Then, the model is reconstructed by using the hidden units to reconstruct the visible units and finally recompute the hidden layer using the reconstructed visible layer. As shown in Figure 2 (left), there are symmetric and undirected connections between the visible and hidden layers. These connections represent the weight matrix or features that need to be learned after the RBM network is converged to the right solution. The learning process is based on minimizing the energy function according to the quality of the reconstructed image and by using the contrastive divergence algorithm. Of course, there are several parameters, like the learning rate, momentum, and weight decay, play important roles in extracting interesting features, so fine tuning these parameters is important to make them appropriate to the given Arabic datasets.

Disadvantages:

- It has been shown that using binary units for the reconstructed visible layer is not appropriate for multivalued inputs like pixel levels of handwritten Arabic letters.
- The multitude of RBM-layers in addition to the size for every RBM-layers (number of units) depends on the final recognition results and the overall classification complexity.

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

3.2 proposed model (epso-eknn)

Figure 3 shows a schematic representation that highlights the major modules of the developed AOCR mechanism. EPSO is used to identify the best features, and EKNN-Classifier is used to recognize Arabic letters in every class. The mechanism is divided into 2 stages: Training and Testing. The AOCR mechanism methodology is divided into various phases. Mostly all OCR processes share the majority of these phases. Feature selection and classification are the two phases examined in depth in this research work.

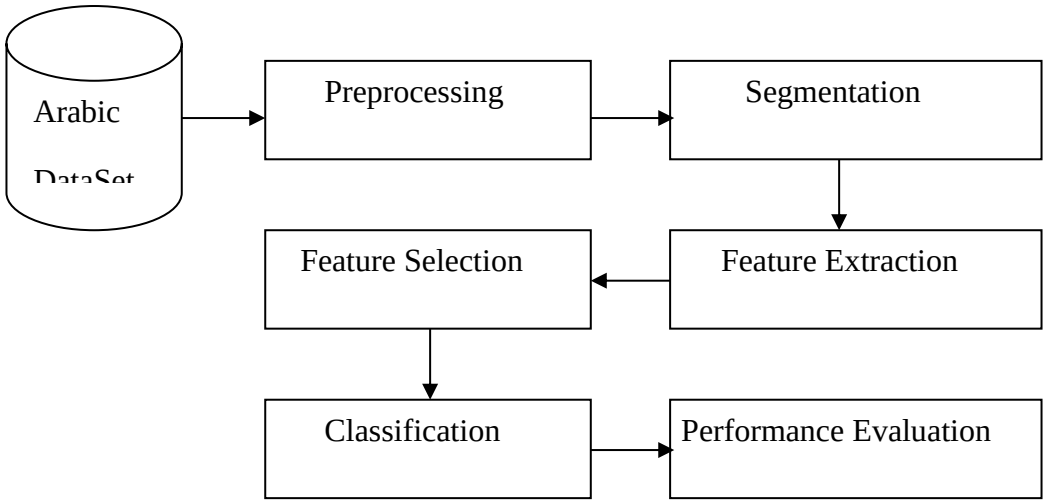

Figure 3: Proposed AOCR Methodology

3.2.1 Arabic dataset

Even though there were numerous standard Arabic datasets, the suggested model relied upon the best printable datasets with excellent quality, a variety of dimensions, and types of fonts. The "**APTI**" dataset, which contains 113,285 textual-images, 12 Arabic-fonts, 11 sizes of fonts, and 5 styles of font, were examined. The dimensions, font styles, alignment, and noise level with different sampling variations.

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

3.2.2 PRE-PROCESSING

Pre-processing is a standard requirement in terms of improving the accuracy of classification. To create normalized and centralized character images, pre-processing processes have been used. OCR performance is significantly improved by this pre-processing. The preceding procedures were used for pre-processing in this research.

(i) Normalization of Sizes:

After getting the character image from the dataset as an input. The image must be normalized and consolidated in an attempt to obtain the highest recognition accuracy. The image dimension is resized to a set dimension by normalizing. Most of the images in this category have been scaled down to 64 by 64 pixels and transformed to grayscale coloring maps.

(ii) Centralizing:

Characters appeared in numerous orientations in many of the images (left, right, bottom and top). To begin, the character's and image's centroid are computed independently in an attempt to align all of the characters around the same place and compute correct characteristics for each Arabic-characters. Because the character is 64 by 64 dimensions in scale, the character's center point is 32 by 32 in this scenario. The character's centroid is therefore moved to the image's centroid to create a centered image.

As illustrated in Figure 4, every sample's image goes through 5 procedures to qualify for segments. The above activities include: (a) converting the images into grayscale and also to binary sequence, (b) attempting to remove noisy data from images using an appropriate Median-Filter, (c) attempting to remove all tiny artifacts using morphology operation of close and open, (d) rotating the image, and (e) reformatting the image to applicable measurements to manage the dimension issue because a few of the characters seem to be smaller.

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

Figure 4: AOCR Preprocessing process

3.2.3 SEGMENTATION

Considering text segmentation seems to be an important contributor to recognizing problems, the presented approach skips these processes and instead relies on pre-segmentation (words free segmentation). Images through the dataset, on the other side, would be split individually since they are lines of text. A Line-Segmenting method was used in this research which identifies and separates individual textual lines from a computerized image text for subsequent extracting features processing after the pre-processing process. The following difficulties must be overcome in the procedure of separating the characters with this Line-Segmenting: (i) Line borders that overlapped, (ii) Lines that contact, (iii) Shattered lines, (iv) Absence of basic data, (v) Curved letter, (vi) Piece wise straight letter, (vii) Connecting letters and phrases inside the lines.

3.2.4 FEATURE EXTRACTION

Following segmenting, the extraction of features stage's primary objective is to maximize detection performance with the fewest amount of features contained in a vector space. The objective of this process is to extract different various features from the character segmented image which have a significant level of resemblance between sampling of the respective classes and a significant level of variance between samples of different classes. Second-order fact based techniques outperformed power spectral density (transformation) and structuring approaches in terms of differentiating levels. Image intervals delivered the greatest outcomes from such second-order facts. As a result, the suggested framework uses a collection of 14 features derived from

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

Gray-Level Co-occurrence-Matrix (GLCM) that seem to be Scale-Invariant and Translation and are therefore reliant on invariance moments.

The first-order facts of a character image, which are focused on specific pixel features and are derived from Standard-Deviation and Mean. GLCM, essentially accounted for the spatially interrelationship or co-occurrence of 2 pixels at certain relative locations, may be used to derive an image's second-order facts. The fourteen features of Angular-Second-Moment, Correlations, Contrasting, Sum of Squares or Variance, Inverse-Difference-Moment, Sum-Average, Sum-Entropy, Sum-Variance, Difference-Entropy, Difference-Variance, Information-Measure of Correlations, and Cluster-Tendency have been determined for the orientations of each matrix. Homogeneity, Entropies, Contrasting, and Energies are all impacted by the orientation chosen. Upon that foundation of the frequencies, the Entropies and Homogeneity provide an indicator of the dominant values of the major diagonal. The energy provides information about the geographical distribution's unpredictability.

The co-occurrence matrices computations have the benefit of allowing co-occurring pairings of pixels to be spatially linked in a multitude of orientations in terms of distance and angular spatial connections, rather than only two pixels at a time. As a result, it seems that the mix of grey levels and their locations will be seen there. The outcome among those feature vectors would then be used in the feature selection procedure.

3.2.5 FEATURE SELECTION (EPSO)

(i) PARTICLE SWARM OPTIMIZATION (PSO)

Eberhart and Kennedy invented PSO in 1995 which was a population related stochastic method. A population with particles has been used to initialize the PSO. Every particle has been seen as a single point inside an "S" dimensional environment. "$X_i = (x_{i1}, x_{i2},..., x_{iS})$" represents the "$i^{th}$" particle. Any particle's previous best position is "p_{best}" (higher value of fitness) was "$P_i = (p_{i1}, p_{i2},..., p_{iS})$". The "$g_{best}$" is the index of the best global particle. The "$V_i = (v_{i1}, v_{i2},..., v_{iS})$" is the particle's velocity of 'i'. These particles are computed using the equations below:

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

$$v_{id} = w*v_{id}+c1*rand()*(p_{id}-x_{id})+c2*Rand()*(p_{ad}-x_{id}) \quad \text{Eq}\rightarrow 1$$

$$x_{id} = x_{id} + v_{id} \quad \text{Eq}\rightarrow 2$$

Here 'w' has been the weight of inertia. The weighting of the stochastic accelerating factors which drive every particle nearer "p_{best}" and "g_{best}" locations is represented by the accelerated constants "c1" and "c2" in Equation (1). The "rand()" and the "Rand()" functions are two randoms with [0,1] routines. The velocities of particles within every dimension were restricted to a "Vmax" maximum-velocity.

This Basic PSO was designed to solve issues involving continual optimizing. The basic PSO idea must be modified to cope with binary information in need to conduct better FS. The Fitness-Function 'f' will be a discrete-function, and the search-space 'D' could be a limited collection of states. The literature proposes many variants of binary and discrete based PSO.

(ii) Extended PSO (EPSO) for Feature Selection:

The position of the particle is represented as "N" length for binary-bit strings, whereby 'N' represents the overall range of attributes. Each bit denotes an attribute, a value of '1' indicates that the associated attribute is chosen, while a value of '0' indicates that the attribute is not selected. In which every position is a subset of an attribute. After calculating velocities and positions using Equations (1) and (2), next it performs a sigmoid-transformation to the velocity component by Equation (3), this compresses velocities mostly with ranges [0, 1].

$$S(v_{id}^{new}) = \frac{1}{1+e^{-v_{id}^{new}}}$$

$$\text{if } (rand < S(v_{id}^{new})) \text{ then } x_{id}^{new} = 1;$$

$$\text{else } x_{id}^{new} = 0 \quad \text{Eq}\rightarrow 3$$

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

Hereby "x_{id}^{new}" has been the current-value of the "i" individual in the "d" dimension and "v_{id}^{new}" has been the current velocity of the "i" individual in the "d" dimension.

Fitness-Function: The accompanying fitness function has been used in this research.

$$\text{Fitness} = \alpha * \gamma(F_i(t)) + \beta * \frac{|N| - |F|}{|N|} \qquad \text{Eq}\rightarrow 4$$

Whereas "$F_i(t)$" is the subset of features discovered by particle 'i' of 't' iteration, "$\gamma(Fi(t))$" has been the quality of the classification for the selected features, the |F| has been the length of the subset of the features selected. The entire feature number is represented by |N|. With "$\alpha \in [0,1]$ and $\beta = 1-\alpha$", 'α' and 'β' were two parameters that relate to the significant quality of classification and the length of the subset length. It determined that the efficiency of the classification is much more relevant than the length of the subset in this research, thus these were fixed to "$\alpha = 0.85$" and "$\beta = 0.15$".

With each iteration of this process, the inertia weight lowers in this method. The size of the swarm had been setted as 30 and the coefficient weight had been setted as 1.2 initially and 0.4 in the final. The accelerating positive constants 'C1' and 'C2' had been setted to 2. Figure 5 depicts the schematic of the proposed FS algorithm.

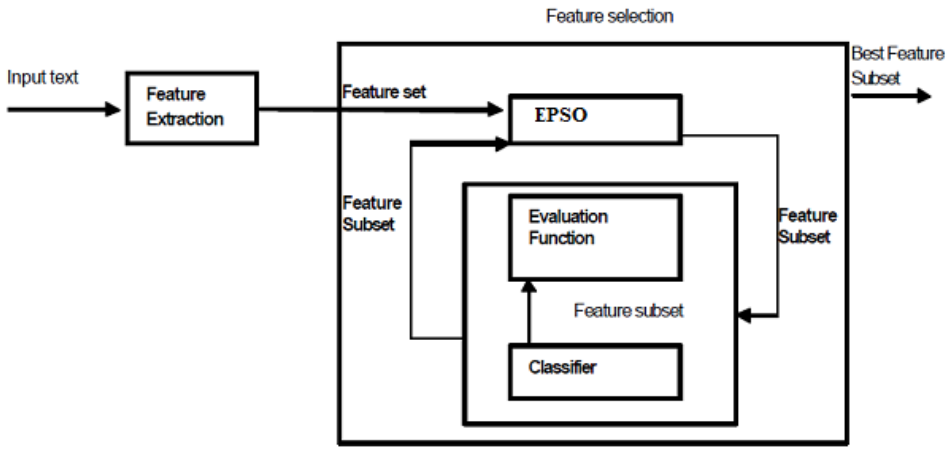

Figure 5: Proposed FS Schematic diagram

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

The EPSO mechanism is implemented in the following way:

- Create a particle's population having randomized position and randomized velocities in the feature-space on 'S' dimensions. "P_i" should be initialized with copies of "X_i", and "P_g" should be initialized with the index of the particle in the population with the greatest fitness-function value.

- Evaluate the required optimization fitness-function as by Equation (4) in 'd' variables for every particle.

- Comparing the fitness assessment of the particle to the particle's "p_{best}". Set the "p_{best}" value to the current value and the "p_{best}" location to the current position in 'd' dimensional space if the current value is better than "p_{best}".

- Comparing fitness assessment to the population's previous overall best. Reset "g_{best}" to the current particle's array index and value if the current value is better than "g_{best}".

- Equations (1) and (2) should be used to adjust the particle's velocities and positions.

- Loopback to (2) until a criterion is satisfied, which is typically a significant level of fitness or a maximal frequency of iterations "generations".

The Arabic text is represented using a vector space representation, and the weight is determined using the Equation below.

$$w_{kj} = \frac{tf \times idf(t_k, d_j)}{maxtf}$$

Eq→5

The weight of the character 'k' in the document 'j' is "w_{kj}". The Term-Frequency is denoted by "tf" (determines the significance of the character in the document). The Inverse-Document-Frequency is denoted by "idf" (determines the significance of character throughout the entire collections).

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

Experimental configuration for FS based on EPSO:

The following are the key stages in the EPSO based FS simulation:

- We've divided the data into 10 groups. In the Arabic-Dataset, each group includes training and testing documents for each category. With a ten percent increase, it had been included negative examples from different categories to each group for both training and testing documents.
- Each group's documents have been preprocessed.
- Then, for each group, the EPSO based FS methodology is utilized to the entire feature space to choose the optimal FS subset that best represents the FS space.
- According to the defined learning method, the optimized FS subset was deployed to a classification algorithm to execute the categorization job (binary classifying).

3.2.6 CLASSIFICATION (EKNN):

(i) KNN

The KNN is being a supervised technique of learning which is one of the simplistic of Machine-Learning methods. The technique uses the closest k-samples from training-sets to classify an unknown sample-class. KNN originally stood for "nearest k samples", which may be found by the distance calculation. The unknown-category is the one with the highest numbers in class.

Figure 6's left side, for instance, shows 3 kinds of samples as "Square", "Circle", and "Triangle" and also an unknown-sample as "Rhombus". The four nearest samples have been chosen. The number represents the distance's order. The closest is 1 and the longest distance is 4. Class-A has a higher probability, as shown in Figure 6. As a result, the unknown-sample is assigned to Class-A.

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

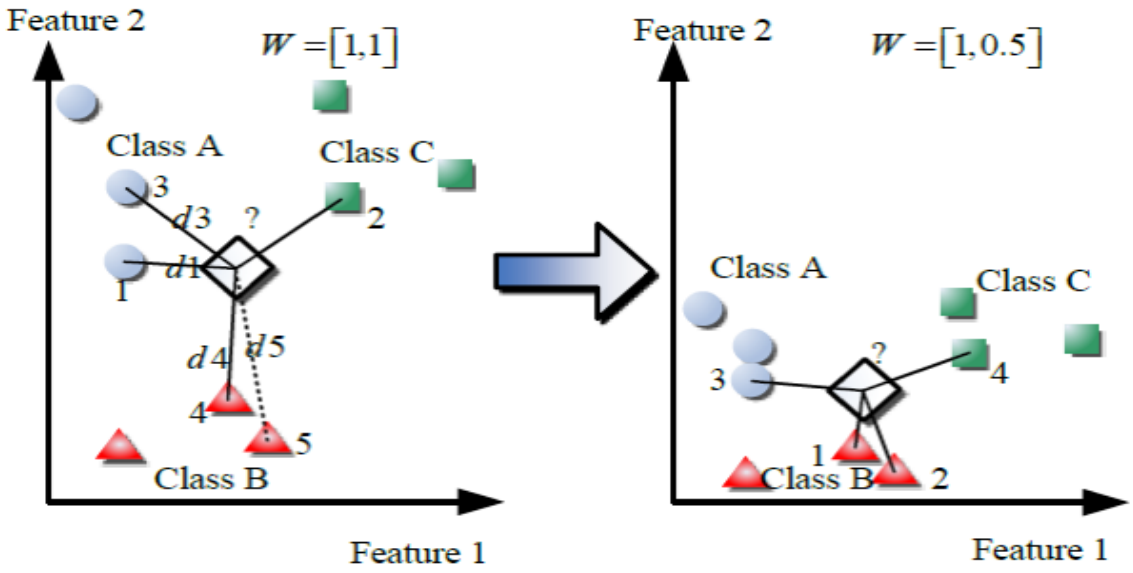

Figure 6: Schematic of Enhanced-KNN (EKNN) Classification

To demonstrate the similarity, K-NN uses a distance metric, typically Euclidean-Distance (ED). If "X = (x1, x2,..., xn) and Y = (y1, y2,..., yn)" are n-dimensional vectors, the ED computes as follows:

$$dist(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

Eq→6

Even so, k-NN has the following shortcomings:

- The weights of all features were equivalent. The positive-correlation does not exist for all features of the categorized outcomes. The error categorized outcomes are due to improper features.

- Identically classified probabilities will likely arise. If a nearest-neighbor is included in the classifying choices, the probability of Class-A, Class-B, and Class-C are 40%, 40%, and 20%, correspondingly, as seen on the left-side of Figure 5. The probability of Class-A and Class-B will be the same.

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

(ii) Enhanced-KNN (EKNN)

The classification outcome of KNN was sensitive to features, which is a KNN drawback. The performance of the classifier is lowered by improper features. The EKNN is used to add weights to overcome this disadvantage. Numerous weights were employed for various features. The weight "W = [w_1,w_2,..., w_n]" is used to calculate the weighted distance in "x" as described in the following:

$$dist(X,Y) = \sqrt{w_1(x_1 - y_1)^2 + w_2(x_2 - y_2)^2 + \cdots + w_n(x_n - y_n)^2}$$

Eq→7

The weight on the left side of the figure is "W = [1, 1]", whereas the weight on the right side of the figure is "W = [1, 0.5]". The above Equation (7) demonstrates how weight changes may affect classifying outcomes. Since Class-A has a higher probability at first, once the weight is adjusted, Class-B will have a higher probability than Class-A. Feature 2 has a much less impact on classifying outcomes. To improve the accuracy of classification, the weight adjustments may change the impact of various features on the recognition rate and remove some incorrect features.

This research study proposes that distance determination be used to establish the classification of the KNN outcome that produces equal probability. Class-A and Class-B have had the same probability in Figure 6's left side. This technique calculates the distance between Class-A and Class-B and each of their KNN's two test-points. The distances between Class-A and its KNN test-points are d1 and d3. The distances between Class-B and its KNN test-sites are d4 and d5. The total of smaller distances to the class is regarded as the predictor in classification if D is the decision outcome.

$$\begin{cases} D = \text{Class A if } d1 + d3 < d4 + d5 \\ D = \text{Class B if } d1 + d3 > d4 + d5 \end{cases}$$

Eq→8

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

This technique would eliminate the same probability and improve classification accuracy. EKNN is a weighted KNN upgraded classification based on EPSO data from FS. With Leave-One-Out-Cross-Validation "LOOCV", the EPSO optimizes weights and k value and evaluates the prediction classification accuracy of EKNN. Every data sample is a class in the LOOCV. Within every calculation, one class would be a testing-sample, while the remaining were training-samples. The total of recognized accuracy is calculated as Predictive-Classification-Accuracy "pcaCV" for all classes. The "pcaCV" is represented as in following Equation (9) whereas if the number of accurate-classifications is "Ncorrectly" and also the number of overall training-samples is "Ntotal":

$$pca_{CV} = \frac{N_{Correctly}}{N_{Total}} \times 100\% \quad \text{Eq}\rightarrow 9$$

IV RESULTS AND DISCUSSIONS

A significant number of tests have been conducted to assess the proposed EPSO-EKNN method. This section includes the comparative results between the DBN and EPSO-EKNN in an attempt to evaluate the effectiveness of this research work. Here it utilized the APTI-Arabic databases for this recognizing the Arabic characters. For the classification work, the code was written and developed in the Matlab2016 platform. Furthermore, it performed the tests on a computer with a 2.90-GHz Processor, with 16-cores, and 16-GB of RAM-Memory. It's important to note that the majority of performance data is presented in both descriptive and analytical formats. Numerous model parameters are evaluated to verify these methods, and the optimum metrics, such as Accuracy, Precision, and Recall, are employed inside these experiments.

(i) Accuracy

The confusion-matrix between both the actual information and also the character recognition (CR) output is used to measure accuracy. The following formula should be used to determine accuracy:

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

"Accuracy = (True-Negative + True-Positive) / (True-Negative + True-Positive + False-Negative + False-Positive)**"**

Arabic-Datasets	DBN	EPSO-EKNN
ArabicImage-1	89.5	94.5
ArabicImage-2	90.5	95.5
ArabicImage-3	89.5	94.5
ArabicImage-4	88.5	93.5
ArabicImage-5	87.5	92.5

Table 1: Numerical Accuracy Comparison

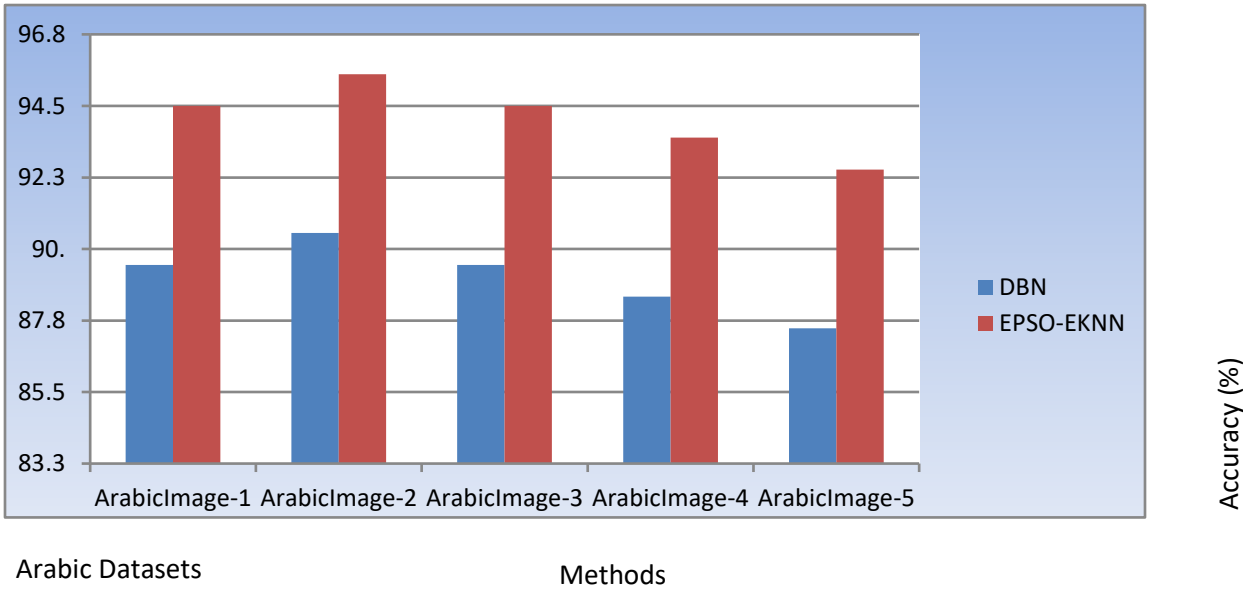

Figure 7: Graphical Accuracy Comparison

This procedure is carried out on all images in the datasets. The findings are produced after evaluation upon CR utilizing actual reality information. Table 1 and Figure 7 shows the maximum accuracy of the CR was lower without selecting optimal features for AOCR by DBN and higher for following optimal feature selection by EPSO for AOCR by EPSO-EKNN.

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

(ii) Precision

The Precision refers to the ratio of the number of character recognition in the particular image that was correctly assigned in that respective category class to the total number of images classified as belonging to respective categories.

"**Precision** = (True-Positive) / (True-Positive + False-Negative)"

Arabic-Datasets	DBN	EPSO-EKNN
ArabicImage-1	90	95
ArabicImage-2	91	96
ArabicImage-3	90	95
ArabicImage-4	89	94
ArabicImage-5	88	93

Table 2: Numerical Precision Comparison

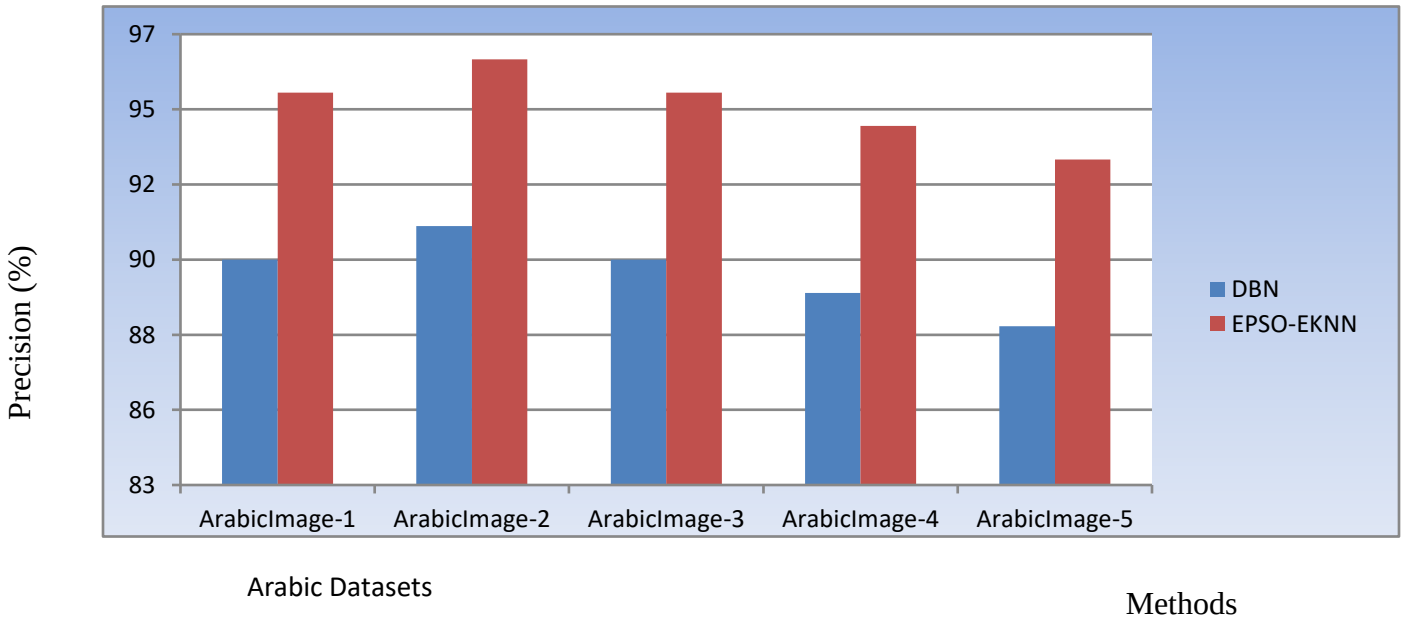

Figure 8: Graphical Precision Comparison

This procedure is carried out on all images in the datasets. The findings are produced after evaluation upon CR utilizing actual reality information. Table 2 and Figure 8 shows the maximum

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

precision of the CR was lower without selecting optimal features for AOCR by DBN and higher for following optimal feature selection by EPSO for AOCR by EPSO-EKNN.

(iii) Recall

Recall refers to the ratio of the number of characters correctly assigned in the respective category to the total number of images belonging to that particular category in original datasets.

"**Recall** = (True-Positive) / (True-Positive + False-Positive)"

Arabic-Datasets	DBN	EPSO-EKNN
ArabicImage-1	91.5	96.5
ArabicImage-2	92.5	97.5
ArabicImage-3	91.5	96.5
ArabicImage-4	90.5	95.5
ArabicImage-5	89.5	94.5

Table 3: Numerical Recall Comparison

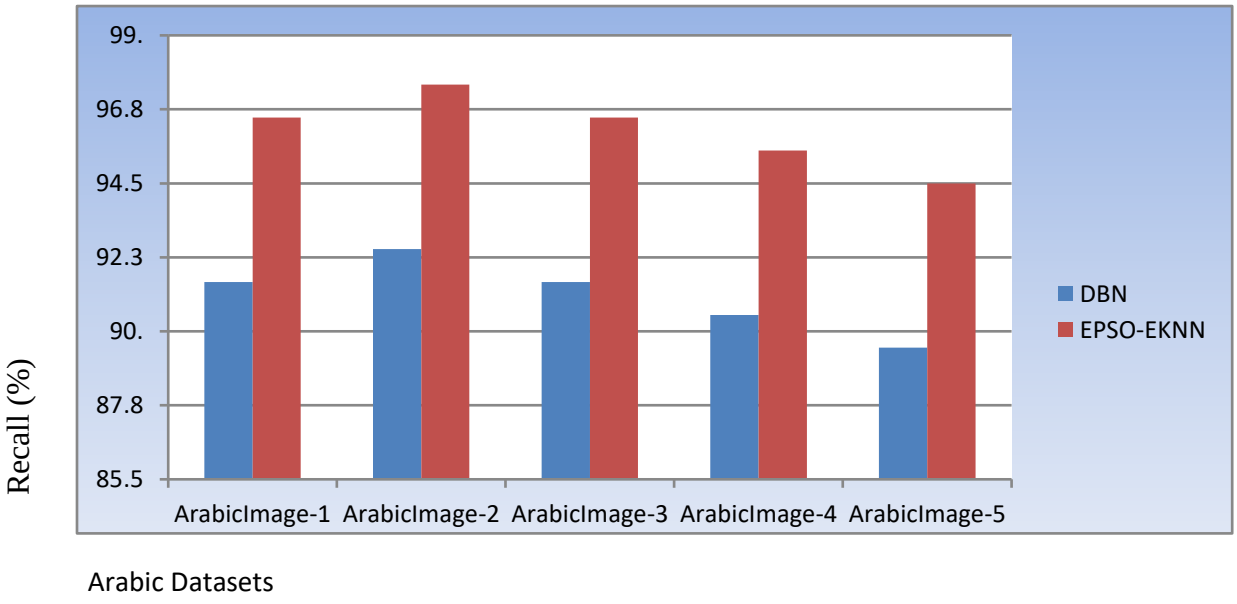

Figure 9: Graphical Recall Comparison

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

This procedure is carried out on all images in the datasets. The findings are produced after evaluation upon CR utilizing actual reality information. Table 3 and Figure 9 show the maximum recall rate of the CR was lower without selecting optimal features for AOCR by DBN and higher for following optimal feature selection by EPSO for AOCR by EPSO-EKNN.

V. CONCLUSION

As perhaps the greatest difficult stage in the OCR research process is the CR. Cursive-Character recognition has sparked a lot of attention in the academic sector over the past couple of decades. The lack of a uniform dataset, on the other hand, renders it very difficult. This article presents an OCR method for Arabic characters to solve these issues. In this paper, we present an extended version of standard PSO (EPSO) for FS and an enhanced version of standard KNN (EKNN) classifier for AOCR Classification, which was tested on the Arabic-Dataset. PSO's performance is influenced by the Inertia-Parameters (w), position-updating method, and fitness-function. We experimentally modified and tweaked PSO parameters in our research. Through all the optimization of the process parameters for PSO, the weighting after modification may impact the features of the EKNN classifier and significantly improve the accuracy of classification. The FS could efficiently reduce classifier calculation time and remove non-optimal features without decreasing the accuracy of classification. The results of the experiments revealed that the created EPSO-EKNN methods outperformed the current DBN methods in aspects of classifiers Accuracy, Precision, and Recall. This research may be facilitated in the future by concentrating on segmentation and classification using sophisticated learning algorithms.

REFERENCES

[1]. Baldominos A, Sa´ez Y, Isasi P (2019) A survey of handwritten character recognition with mnist and emnist. Appl Sci 2019:3169

[2]. Hasanuzzaman, H., 2013. Arabic language: characteristics and importance. J. Humanities Soc. Sci. 1 (3).

[3]. Abdalkafor AS (2018) Survey for databases on Arabic off-line handwritten characters recognition system. In: 2018 1st International conference on computer applications information security (ICCAIS), pp 1–6

ISSN: 0011-9342 | Year 2021
Issue: 8 | Pages: 4728-4750

[4]. Ahmad I, Fink GA (2016) Class-based contextual modeling for handwritten Arabic text recognition. In: 2016 15th International conference on frontiers in handwriting recognition (ICFHR), pp 554–559

[5]. Mansour, A.M., Hawashin, B., Aljawarneh, S., (2013). An Efficient Feature Selection Method for Arabic Text Classification. Int. J. Comput. Appl. 83 (17), 1–6.

[6]. Azuraliza, A.B., Siti Rohaidah, A., Nurhafizah Moziyana, M.Y., Yaakub, M.R., (2017).Statistical analysis for validating ACO-KNN algorithm as feature selection in sentiment analysis. International Conference on Electronics and Communication System.

[7]. Maalej R, Kherallah M (2018) Convolutional neural network and blstm for offline Arabic handwriting recognition. In: 2018 International Arab conference on information technology (ACIT), Werdanye, Lebanon, 2018, pp 1–6.

[8]. S. Khan, A. Hafeez, H. Ali, S. Nazir, and A. Hussain, (2020)“Pioneer dataset and recognition of handwritten Pashto characters using convolution neural networks,” Measurement and Control, vol. 53, no. 2, p. 20294020964826, 2020.

[9] S. Khan, H. Ali, Z. Ullah, N. Minallah, S. Maqsood, and A. Hafeez,(2019)“ KNN and ANN-based recognition of handwritten pashto letters using zoning features,” 2019, https://arxiv.org/abs/1903.10921.

[10]. Alghamdi, H., Selamat, A.,(2014). The hybrid feature selection k-means method for arabic webpage classification. Jurnal Teknologi. 70 (5), 73–79.

[11]. Azuraliza, A.B., Siti Rohaidah, A., Nurhafizah Moziyana, M.Y., Yaakub, M.R.,(2017). Statistical analysis for validating ACO-KNN algorithm as feature selection in sentiment analysis. International Conference on Electronics and Communication System.

[12]. Mudhsh MA, Almodfer R (2017) Arabic handwritten alphanumeric character recognition using very deep neural network. Information 8(3)

[13]. Younis K (2018) Arabic handwritten characters recognition based on deep convolutional neural networks. Jordan J Comput Inform Technol (JJCIT)

[14]. Peng, C., Limc, S., Chin Neoh, L., Zhang, K., Mistry, K., (2018). Feature selection using firefly optimization for classification and regression models. Decis. Support Syst. 106, 64–85.

[15]. Hasasneh, Ahmad & Salman, Nael & Eleyan, Derar. (2019). Towards Offline Arabic Handwritten Character Recognition Based on Unsupervised Machine Learning Methods: A Perspective Study. 1. 1-8.