How to Cite:
Ashiq, V. M., & Fredrik, E. J. T. (2022). An OCR for Arabic character recognition with advanced principal component analysis based on feature extraction and fuzzy-KNN based classification. *International Journal of Health Sciences*, *6*(S1), 12205–12224.
https://doi.org/10.53730/ijhs.v6nS1.7918

An OCR for Arabic character recognition with advanced principal component analysis based on feature extraction and fuzzy-KNN based classification

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

Dr E J Thomson Fredrik
Professor, Dept. of Computer Applications, Karpagam Academy of Higher Education, Coimbatore

***Abstract*---**Offline character recognition has become a highly important study field for various pattern recognition applications in recent years. Several handwritten character recognition systems have been suggested, with the complexity of these systems varying depending on the recognizing units' writing styles. In reality, identifying letters or numerals is far simpler than recognizing cursive sentences or lines of text. As a result, early handwriting recognition algorithms could only distinguish a few characters with limited vocabularies. Nowadays Arabic handwritten character recognition is very important as it is very difficult to identify. The cursive writing and variety of styles make this recognition more complex. This paper presents an automated model for ACR. This ACR is constructed from four phases: Preprocessing, Segmentation, Feature Extraction, and Classification. In this research article, we compare the advanced EPSO EKNN Algorithm with the earlier DBN Algorithm and also suggest a new method having higher Accuracy. In this research article, the "Enhanced K-Nearest Neighbor" (EKNN) classification model was used to identify and classify or simply recognize the particular Arabic character, and the "Extended Particle Swarm Optimization" (EPSO) methodology was introduced to select the best feature from feature extraction to comply with Arabic-Character Classification. As compared to Deep belief networks, the developed EPSO –EKNN Algorithm has higher accuracy.

***Keywords*---**ACR, Fuzzy, KNN, Principal Component Analysis.

Introduction

There are several uses for OCR, like documentation recovery, postcode and vehicle license plate identification, and many more financial and commercial operations [1]. Online-OCR and Offline-OCR systems are the two main types of OCR. Handwritten content may be recognized using Online-OCR because it detects characters as they have been typed, using the pace, orientation, and sequence of separate brushstrokes to reach elevated levels of detection performance. Offline-OCR, on the other hand, is more complicated [2]. Many challenges must be overcome in terms of achieving this level of identification, including the similarity of various character forms, the links between nearby characters, and the overlapping between characters. As a result, Offline-OCR tools are widely employed in activities like translating handwritten mailing information on the package and analyzing banking cheques and cash deposits [3].

As a result of Offline-OCR is capable to rewrite ancient and heritage records in digital form, it reduces time and manpower [4]. Research into Offline-OCR is exciting because of the challenges it faces, and also the growing demand for OCR application areas [5]. The OCR seeks to achieve a higher identification percentage, overcoming the inferior quality of digitized imagery, especially in ancient records, and adjust to size and style differences within one single document [6]. Since Arabic has such a complex word pattern and grammar, AOCR is continually evolving, irrespective of different languages [7].

The following are a few of the challenges:

- As illustrated in Table 1, each character possesses 2 or 4 distinct forms. The shape of every letter is determined by its position in the world world.
- The design of certain letters is identical, but the location and dots count of characters including such (ث ت ب), that may be placed anywhere below or above the characters, might vary greatly.
- The characters are interconnected. These "Pieces of Arabic Words (PAWs)" are characters that could indeed be accompanied by the later characters which make a word comprising several linked components.
  This is further demonstrated in Figure 1 through the usage of punctuation marks (special markings placed below or above the character).

Table 1
Forms of Arabic-Characters

Name	Isolated	Initial	Medial	Final
Alif	أ	---------		ا
Baa	ب	بم	بم	ب
Taa	ت	تح	تح	ت
Thaa	ث	ثن	ثن	ث
Jiim	ج	جم	جم	مج
Haa	ح	حم	حم	مح
Khaa	خ	خم	خم	مخ
Daal	د			د
Zaal	ذ	---------		ذ

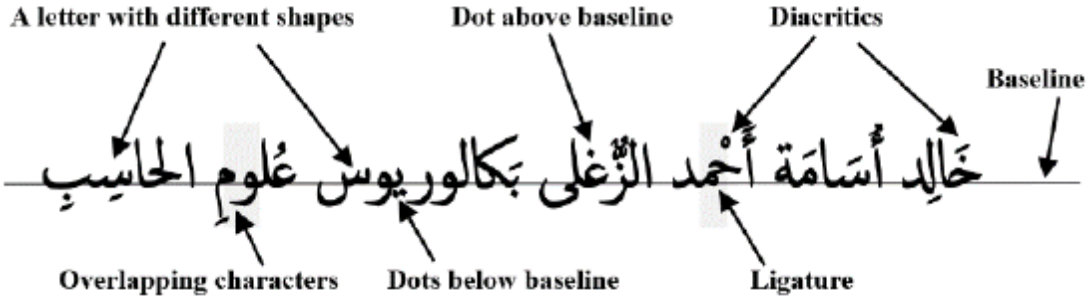

Figure 1: Characteristics of a Printed Arabic-Script

The research's problem statement focuses on ways to receive extremely excellent AOCR which tackles the odd structure of Arabic-Characters while achieving significant detection accuracy and coping with a wide variety of fonts and sizes [8]. For AOCR classification, the Enhanced version of the K-Nearest-Neighbor (EKNN) algorithm was earlier implemented. The most of EKNNs in a source sample groceries assigned to the object 'x' in EKNN. With each class, a rough approximation of the posterior probability is provided by a fraction of the f neighbor chosen for that particular class. When it comes to complex characters of the same type, this fails in computing precision and accuracy [9].

The main contribution of this research work concentrates on feature extraction and classification phases. For feature extraction in this research work, we had implemented an Advanced version of the Principal Component Analysis (APCA) for extracting features based on the optimized transformation to extract more optimal features. For classification in this research work, we had implemented Fuzzy-KNN. Excited by the prospect of "fuzzifying" traditional classifications, fuzzy classification models have been developed. Whenever they are available, FKNN utilizes the distances to neighbors and also their Soft-Labeling. Rather than allocating a pattern to a single class, FKNN uses a Fuzzy-Set to determine whether or not it belongs in a certain category. It's important to bear in mind that Fuzzy-Systems may be quite effective since they aren't too sensitive to things like

shifting settings or mistaken or missed rules. When it comes to processing power, the working procedure is typically simple, unlike the exact computational models, which save resources. It's an intriguing characteristic, especially for practical applications. The model's goal is to achieve the lowest identification error, the fastest runtime, and the smallest structure possible.

The remaining sections of this research article are organized by the following sections: Section discusses some recent articles related to the problem of OCR, Section 3 details the proposed methodologies module by module also with crisp details of an existing method, Sect and n 4 show the results and comparison obtained for both existing and proposed methods with different parameters and finally, Section 5 concludes this research article.

Related Works

Mars A, Antoniadis G (2016)[10] created an Arabic database with "6090-Characters" and "1080-Words" for word and character recognition. The classification was done by employing a "Neural Network" and a "Time-Delay Neural Network (TDNN)". It has a letter accuracy of 98.50 percent and a word accuracy of 96.90 percent.

A fresh dataset including "16,800 Arabic Characters" was released by El-Sawy A, Loey M, Hazem E (2017) [11]. The "Center for Microprocessor Applications for Training, Education, and Research Database CMATERDB 3.3.1'" Handwritten Arabic digits database includes "30distinct sample sales with "32*32 pixels RGB bitmap images", providing 3000 individual data. At "16,8000-Characters long", the AHCD database includes the writings of 60 individuals, who are all between the ages of 19 and 40.

Handwriting recognition using a combined "CNN-BLSTM" framework developed by Maalej R, Kherallah M (2018) [12]. To automatically retrieve features from input data, CNN has been used. The "Classification Temporal Layer" is connected to the "Bidirectional Long Short-Term Memory (BLSTM)" and is employed for sequential labeling. The combination model's recognition accuracy is 92.21 percent, according to tests done on the "IFN/ENIT" Dataset.

Using a CNN model, Latif G, Alghazo J, Alzubaidi L, Naseer MM, Alghazo Y (2018) [13] were able to identify handwritten composite digits in languages including certain Urdu, Devanagari, Persian, Western Arabic, and Eastern Arabic. The input layer is "28*28 pixels" in size, accompanied by 2 hidden Convolution-Layers with a "5*5 kernel" window size Max pool pool-Layers by including a kernel of 2*2 followed the Convolution-Layers. The integrated Multilanguage dataset has a total accuracy of 99.26 percent and an absolute precision of 99.29 percent. Each language's accuracy was determined to be 99.322 percent on average.

]. F. Alotaibi, M. T. Abdullah, R. B. H. Abdullah, R. W. B. O. K. Rahmat, I. A. T. Hashem, and A. K. Sangaiah,(2018) [14] developed a method for classifying features in the AOCR using a "Neural-Network". Tokens representing the identities of the characters are generated by this method. The proposed technique relies mostly on obtaining a collection of features for each character. Finally, the

identification and building processes use this information. The overall recognition accuracy, on the other hand, is just 87 percent.

Methodologies

Figure 2 shows a schematic representation that highlights the major modules of the developed AOCR mechanism. Here APCA is used to extract the best optimal features, and Fuzzy-KNN is used as a Classifier to recognize Arabic-Characters 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 Extraction and Classification are the two phases examined in depth in this research work.

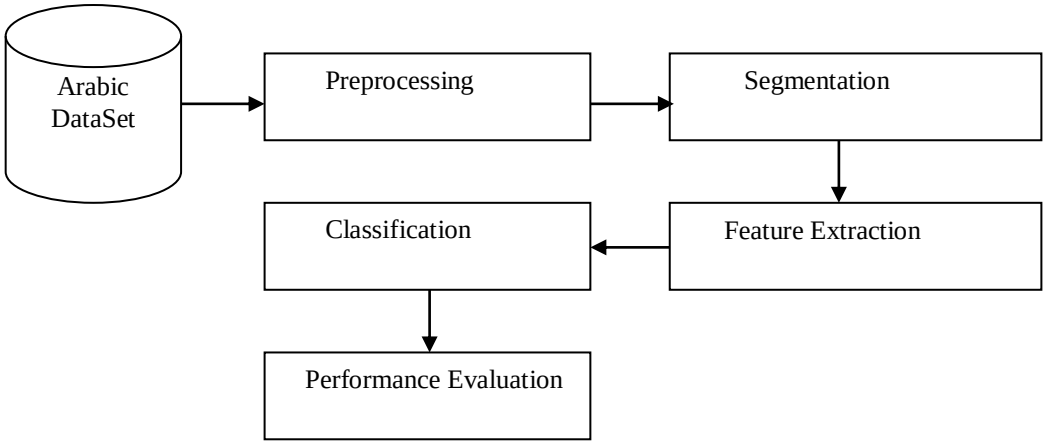

Figure 2: Proposed AOCR Methodology

Existing Model

This method presents an Extended version of the standard "Particle Swarm Optimization" PSO (EPSO) for "Feature-Selection" FS and an Enhanced version of the standard KNN (EKNN) classifier for AOCR Classification, which was tested on the Arabic Dataset [15]. PSO's performance is influenced by the Inertia-Parameters (w), position-updating method, and fitness function. It experimentally modified and tweaked PSO parameters in this method. 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. 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 FS could efficiently reduce classifier calculation time and remove non-optimal features without decreasing the accuracy of classification.

Proposed Model

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.

Preprocessing

The Histogram-Equalization (HE) approach was applied in this experimentation to preprocess the Arabic-Character database. We employ the HE approach to improve the quality of most of the images that have been used in this research and emphasis the text for the segmentation and classification phases because of the inferior quality of a few of the images. This database contained a few low-quality images. It is difficult to see the text, there is a lot of noise, and the backgrounds are not crisp. There are very few drawbacks to using the HE approach, based on our examination and review of the literature available. We modify an earlier adaptable HE approach to match our needs. The HE approach is one of the very often used methods for improving image brightness due to its convenience and effectiveness. The method's "Cumulative-Distribution" procedure is used to standardize the "Intensity-Distribution". The "Intensity-Distributions" of the generated images will also be similar. We hope this pre-processing will assist to identify a segmentation approach that will allow us to easily separate text from the background.

Segmentation

The Image-Binarization (IB) approach was employed in this research to segment the Arabic-Character following preprocessing. For the IB, it is necessary to separate a single word or phrase from a larger context. Rather than the standard threshold technique, we apply a model that can be trained. An IB model based on structural edges and texture is capable of capturing text with weak structures and a range of colors. Training and Testing are two separate steps in this process. First, a certain volume of images is chosen for the training process. We deploy about 100 images in this research. Those images have a range of texts, each of which presents a unique challenge. Numerous samples have been taken from the text and also the background of each training image, including a small window of square sections of a certain dimension "e.g. 5*5". Two classes of windows are taken from each Arabic-Character image, with 50 showing Arabic characters (such as lines) and the rest showing a background (including such white noise or undesired items). False-Positive items, including noises, undesired objects, and distortions, may be eliminated with this option. Rather than using image pixels, texture characteristics are used for all windows, regardless of class. As a result, we need to employ a threshold that may assist us to discriminate between Arabic-Character lines and undesired objects by examining each line throughout the image.

Feature Extraction

The PCA is a technique for reducing the dimension of a collection of data while maintaining its heterogeneity. Almost every collection of segmented Arabic-Character includes information expressed as vectors of single parameters with integer, binary, or real values in most cases. A geometric point in three-dimensional space, e.g., may be described by a vector of three parameters each of which is aligned with one of the three coordinates axis 'x', 'y', and 'z'. A segmented Arabic Character could be described in particular by a vector made up of a set of parameters. The length of the vectors found in the collection, and therefore the set's dimension, is determined by the number of parameters. Furthermore, a continuum of variability may be specified for each parameter, that defines the range of values that only the individual parameter could accept. For example, if the data set includes three-dimensional points delimited by a cube with side 1 and centered in (0, 0, 0), the three parameters describing Cartesian coordinates were bound to [-1/2, 1/2]. The spectrum of heterogeneity of the three parameters is described by this interval. The objective of PCA is always to uncover hidden patterns in data and turn them in just such a manner that their differences and similarities are exposed. When the patterns have been discovered, the data may be interpreted as components that are sorted by importance, allowing low-level components to be discarded without losing valuable details.

Advanced PCA (APCA)

When the actual parameters were chosen to describe the segmented Arabic-Character that are associated, the traditional PCA will minimize the dimension of a collection of data. Now let’s reconsider the illustration of the three parameters describing the three positions of Arabic segmented character points inside a cube by APCA. The three Arabic-Character parameters are associated if, for example, any of the points in the collection fall on an appropriate plane.

As APCA has been used to solve this issue, one of the three parameters of specific Arabic-Character is transformed into a void parameter. The points in the fresh transformed space could consequently be defined by just two parameters, resulting in a space with a smaller dimension than the first. Since the points are in a two-dimensional region, the details about the third dimension, which is the rejected dimension, are meaningless. This is an oversimplification of the case. The explanations that follow go into the APCA process in greater depth.

Imply that perhaps the segmented Arabic-Character under consideration includes points in a two-dimensional region with coordinates of (-2, -1), (-1, 0), (0, 1), (1, 2), (2, 3). The values of 'x' differ in the range [-2, 2], whereas the values of 'y' vary in the range [-1, 3]. The variation of the parameters 'x' and 'y' is described by these two intervals. These two factors are associated, as can be shown. As the 'y' coordinates rise, the 'x' coordinates rise as well, and a straight line runs between them all. As a result, if one of the two coordinates is identified, the other may be obtained.

Whereas traditional PCA aims to convert certain parameters so that they are no longer associated. By achieving this, the dimension of the collection of data could

be minimized by only considering the parameters with the highest uncertainty and discarding the others. Principal-Components (PC) are still the factors with the most variability. Here in APCA, the first PC of the Arabic-Character data could have been considered to reflect the results since they are normally ordered by their heterogeneity. The fact that perhaps a lower order PC exhibits lower variance within the ensemble however does not mean this is irrelevant in regression models. Through APCA it is possible to find out the extra Arabic-Character features that are omitted by the traditional PCA.

Extraction of Arabic Character Statistical Features through APCA

APCA is most commonly used for locating a lower-dimensional representation of Arabic Character data. That has 2 distinguishing characteristics. During computing, it initially keeps shrinking the measurements of the segmented Arabic-Character data to a rational and accurate scale. Foremost, it separates the number of distinguishing character features from the segmented Arabic-Character input in a quiet manner that the overall dimension is reduced. The important feature characteristics were always present and could be used to identify the actual Arabic-Character input details.

The covariance matrix could be found again from the matrix by leveraging the collection of optimal Arabic-Character features. The Eigen-values are then calculated using this covariance matrix. Eigen-vectors were effective in representing complete Arabic-Character databases in their nature. Merely some small Eigen-values were considered toward being substantially preferable and greater in importance, whereas the others are substantially quite minimal, and their exposure to data variations is indeed quite limited. As a result, after computing the inner product of the Arabic-Character data well with respective Eigen-vectors for its respective Eigen-values, the preferred and greater variance paths were simply maintained in this proposed APCA.

The following are the general measures of APCA methodology:

Step-1: Compute covariance matrix mix of the specified Arabic Character input from segmented Arabic characters using the following Equation 1:

$$\sum V = \frac{1}{Num}\left\{\left(diag(m) - \overline{diag}(m)\right)\left(diag(n) - \overline{diag}(n)\right)^{T}\right\} \qquad \textbf{Eq}\rightarrow\textbf{1}$$

Where, 1≤ m, n ≤ Num

Step-2: As a result of the Eigen-vector matrix (V) and diagonal-matrix (D) of computed Arabic-Character Eigen-values are:

$$V^{-1}\sum V = D \qquad \textbf{Eq}\rightarrow\textbf{2}$$

Step-3: To achieve the PC parameter, organize the Eigen-vectors in decreasing order with the accompanying magnitude of Arabic-Character Eigen-values.
Step-4: At last, Arabic-Character data is transformed into the form of PCs by measuring the inner product of data with relevant Eigen-vectors.

Through specific, the APCA of a given vector ‘v’ associated with the group ‘V’ is achieved by mapping vector 'v' onto the subspaces with the length or gaps of corresponding Eigen-vectors that relate to the top Eigen-values of the auto-correlation matrix ‘R’ in downwards sequence, where he is lower than e. The above transformation generates a vector of coefficients c1,..., she’s Even so, a linear structure of the Eigen-vectors with its corresponding weights c1,..., he’s' is used to describe the vector 'v'. Thus the statistical features from the segment Arabic Character are derived efficiently by our proposed APCA method. Using such techniques, we may determine the maximum values of each character and the correlation of each phase to which they belong. This APCA process extracts the best optimal features so there is no need for an extra feature selection process. In this way, the overall computation time is get reduced.

Classification

The methods for Classification are primarily employed towards some correlation calculation of a group of items dependent on such distance measurements. One of the earliest classification techniques without dimensionality reduction includes the K-Nearest Neighbor (KNN) method. With Object Membership Utilities, the decision rule for typical supervised learning assumes equivalent weight, ignoring numerous similarity patterns. To improve the Arabic-Character classification rate in this research we propose Fuzzy-KNN (FKNN) approach. The FKNN has been shown that it has not only a lower error in the classification of Arabic-Character but also more faith in the classification taking advantage of the Fuzzy Logic principle. The FKNN gives a more practical vector for the object's membership and thus provides for the object's class membership. A class with its nearest K-neighbors is allocated in this algorithm to the most common form. FKNN assigns the sample's fuzzy membership and allows policymakers to make fuzzy choices. In this research for identifying the specific character from the extracted Arabic-Character features the FKNN was used for classification.

Fuzzy Logic (FL)

The FL identified the role that ties the reality of the proposal to other proposals. A collection of an infinite series, an FL determines the actual [0, 1] number with the number between true and false. Another real meaning is contained in the FL describing the truth tables of varying adjective degrees. The principle presents knowledge linguistically that offers a systemic calculus and linguistic labels indicate numerical estimation utilizing the membership function. A systemic calculus is given. For instance, take into account a classical set A with a narrow limit and contains a real number higher than K which was described as per following Equation 3:

$$A = \{x > K\} \qquad \textbf{Eq→3}$$

It is observed that 'K' is simple and unequivocal. If 'x' is larger than 'K', 'x' will be set 'A', otherwise, it is not set 'A' as in the above equation. In numerous configurations including certain science and engineering, the classical set is used

and does not represent human existence, however, it appears to be complex and imprecise. The fuzzy set simplifies it.

Fuzzy's Terminologies and Basic terms

When 'x' is an 'X' variable, it is object space. A classic-set A is a series of elements such as "such $A \subseteq X$" where "$x \in X$", through 'X' is set 'A' or x is set 'A'. The feature function for any element in 'X' is defined by a classical-set 'A' which defines "$x \notin X$" with a set of ordered pairs (x, 0) or (x, 1). It is allowed to have a characteristic function between 0 and 1 of the fuzzy group, which corral with ates the membership rating of an element within a given set.

Membership functions and Fuzzy-set:

The object selection in 'X' is normally 'x', then a fuzzy-set 'A' in 'X' is represented by a set of pairs organized as in the following Equation 4:

$$A = \{(x, \mu_A(x) : x \in X)\}$$

Eq→4

Membership functioning of the fuzzy set 'A' is indicated in the above equation by "μA(x)". The purpose of the membership function is to map the variable in 'X' from 0 to 1. The fuzzy-set 'A' is described in the following alternative:

$$A = \begin{cases} \sum_{x_i \in X} \mu_A(x_i)/x_i \ \text{if collection of discrete objects in } X \\ \int_X \mu_A(x)/x \ , \text{if continuous space in } X \end{cases}$$

Eq→5

Operation of Set-Theoretic

In traditional sets, complement, intersection, and union are some fundamental features. The fuzzy set is identical to the regular complement, intersection, and union operation. Set 'A' is a subset of 'B' and it's described as "A⊆ B" in a classical approach. Set 'A' and 'B' are both fuzzy-set and subset 'B is 'A' if the following equation only includes "μA(x)≤μB(x)":

$$A \subseteq B \Leftrightarrow \mu_A(x) \leq \mu_B(x)$$

Eq→6

The following equation indicates that 'B' is a sub-set of 'A':

$$A \supseteq B \Leftrightarrow \mu_A(x) \geq \mu_B(x)$$

Eq→7

"C1" and "C2" constitute the union and intersection of two 'A' and 'B' fuzzy pairs were defined by:

$$C_1 = A \cup B \; ; \; C_2 = A \cap B$$

Eq→8

The above correlation's membership-function was derived as:

$$\mu_{C_1}(x) = max\big(\mu_A(x), \mu_B(x)\big) = \mu_A(x) \vee \mu_B(x)$$

$$\mu_{C_2}(x) = min\big(\mu_A(x), \mu_B(x)\big) = \mu_A(x) \wedge \mu_B(x)$$

Eq→9

The '$\bar{A}$' function determines the add-on of the fuzzy-set 'A' as follows:

$$\mu_{\bar{A}}(x) = 1 - \mu_A(x)$$

Eq→10

The next equation describes the well-known feature and operation of the model as:

$$C_W(a) = (1 - a^W)^{\frac{1}{w}} \; and \; \mu_{\bar{A}}(x) = (1 - \mu_A(x)^w)^{\frac{1}{w}}$$

Eq→11

The "$A \times B$" function describes the Product of Cartesian in which 'A' and 'B' is a fuzzy-set in the product field 'X' to 'Y' through the membership-function being:

$$\mu_{A\times B}(x, y) = min\big(\mu_A(x), \mu_B(y)\big)$$

Eq→12

Classification of arabic characters using fuzzy-knn

FKNN is intended to divide the sub-set "X = {x1, x2, . . ., xn}< Rn" of vector samples into the cluster of sub-sets with fuzzy of "c (1<c<n)". In case "(i = 1, 2,. ., c) & (j = 1, 2,. ., n)" the fuzzy matrix of membership is 'U', in which "U_{ij}" is the "xj" fuzzy in class 'i'. The object "j^{th}" is allocated to the class "ith", which has the highest "U_{ij}", relative to the membership of the fuzzy with other groups in a non-fuzzy variant of the method. Two restrictions are present in matrix 'U' as given below:

$$\sum_{i=1}^{c} u_{ij} = 1, \; j$$

Eq→13

$$u_{ij}\hat{I}\,[0,1], \quad 0<\sum_{j=1}^{n}u_{ij}<n.$$

Eq→14

The first-ever limitation as in Equation (12) guarantees that all membership object's grades are earned in all groups (i=1, 2, .,c), and all the membership grades are summed as one. Equation (13) notes that for all objects, the membership of fuzzy classes lies at or above zero and is equivalent to or below one. These 2 limitations illustrate that if an item is in the "U = 1" Arabic Character class, it certainly has no participation in the other Arabic-Character classes. Furthermore, for all Arabic-Character in a class the total of all the membership of fuzzy grades surpasses zero, or else the Arabic-Character class does not exist and is consequently fewer than the total of Arabic-Character 'n'. In the algorithm of FKNN each vector is given the degree of fuzzy membership, regarding the distances between vectors and their membership in KNN which was given below:

$$u_i(x) = \frac{\sum_{j=1}^{K}\left(\frac{u_{ij}}{||x-x_j||^{\frac{2}{(m-1)}}}\right)}{\sum_{j=1}^{K}\left(\frac{1}{||x-x_j||^{\frac{2}{(m-1)}}}\right)}$$

Eq→15

Here 'K' is the neighbors that are nearest with a predefined number, and 'm' is the parameter with a constant. In the measurement of membership of fuzzy value, the 'm' parameter defines the weight of each closest neighbor. In Equation (15) it's a central part in the calculation of the degree of membership of an "i^{th}" object class that was regulated by inverted distances between an 'x' object and its closest neighbors and memberships of a 'K' class.

The reverse association between membership level and distances, unlike most of the non-fuzzy version classification techniques, acts as the part that a weighting feature performs in rewarding/penalizing those with farther or closer distance from other class objects. Clearer is that since an Arabic Character belongs to the 'A' class with a degree of 0.95 when it belongs only to the 'B' class with a degree of 0.05 then it will be rational that an Arabic Character should belong to 'A'.

If therefore the Arabic-Character membership ratings were 0.55 and 0.45 respectively for classes 'A' and 'B', there could be some hesitations in either of the classes 'A' or 'B' until the Arabic Character was allocated. Ultimately, the mission, which creates a greater degree of resemblance, specifies the components of either 'A' or 'B' of the item. The weighting function in Equation (15) illustrates all such circumstances dependent on the inverse distances of the objection set in the class.

The 'm' parameter shows the magnitude when the Arabic Character is assigned/omitted with distances from some of the other Arabic-Character. The near neighbors play a far more significant function in determining the membership standard of the subject to be listed, 'm' is often larger than one and the closer its worth. Through raising this metric, from the other side, neighbors are weighed more equally and are less likely to have relative distances from the categorized Arabic-Character. The distance of Euclidean among its "j^{th}" and 'x' is "$||x-x_j||$"

Whenever the number of input parameters will be less than the scale of the training set, in machine learning models, there is rather a slight chance of Over-Fitting (OF). Cross-Validation (CV) is also implemented for model precision without depending upon the data used during the training set in the calculation of the performance variables.

The CV will focus on saving predictive models from problems like OF and test the independence of the model in its data collection. The OF is rendered by storing the mapping feature from input to output variable when the forecast loses its true sense throughout the process of the testing. This is normal when the data number to be separated between the training and test dataset is not adequate without applying data damage.

Even though the amount of data available is substantially larger than the proportion of input variables, there is a slight possibility that OF is required to obtain an understanding of how quantitative models can be confident, in reality, the Nash-Sutcliffe (NS) process efficiency coefficient can be used. The coefficient of NS is as follows:

$$E_{NS} = 1 - \frac{\sum_{i=1}^{n} (h_i^* - h_i)^2}{\sum_{i=1}^{n} (h_i^* - H)^2}$$

Eq→16

The coefficient for NS ranges from $-\infty$ to 1, in this near to the value of one implies greater preciseness, and an appropriately zero value indicates that the predictive model is a reasonable mean, and some negative values reflect that even the mean of the observed results is less reliable than the predictive model result.

Figure 3 shows the flowchart of the proposed FKNN classifier for classifying Arabic-Character. The FKNN method operates with the number of the closest neighbors by giving an 'x' sample of input and neighbors that are nearest 'K'. With it the first class (i=1) the method is initiated then the distances between 'x' and 'x_i' are determined.

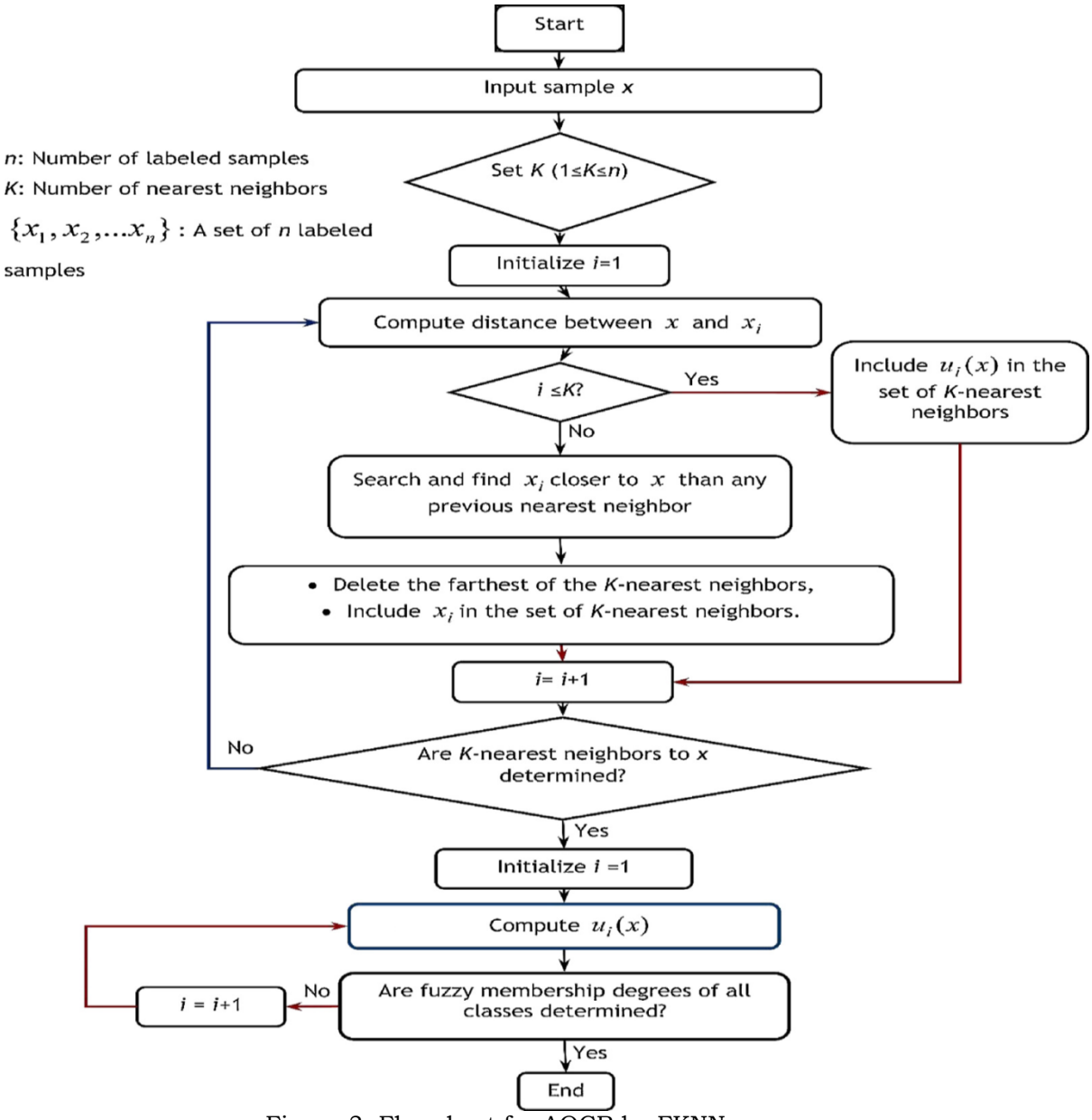

Figure 3: Flowchart for AOCR by FKNN

Moreover, unless the quantity of classes, 'i' is equivalent to or lower than K, the algorithm scans the "x_i" closest to 'X' than the nearest neighbor. In addition to this, "Ui(x)" is included in the K neighboring set. The closest neighbors from K are excluded from the collection if the second condition is valid and substituted by "x_i". The 'i' is then incremented by one and the process is performed until all K-nearest neighbors to the 'x' are found, if the closest 'K' neighbors to the 'x' are not determined. The membership of fuzzy values for both groups is then determined as in Equation (15).

Results and Discussions

A significant number of tests have been conducted to assess the proposed APCA-FKNN method. This section includes the comparative results between the EPSO-EKNN and APCA-FKNN 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.

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

Table 2
Numerical Accuracy Comparision

Arabic-Datasets	EPSO-EKNN	APCA-FKNN
ArabicImage-1	94.5	96.5
ArabicImage-2	95.5	97.5
ArabicImage-3	94.5	96.5
ArabicImage-4	93.5	95.5
ArabicImage-5	92.5	94.5

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 4shows the maximum accuracy of the CR was lower for AOCR by EPSO-EKNN and higher for AOCR by APCA-FKNN.

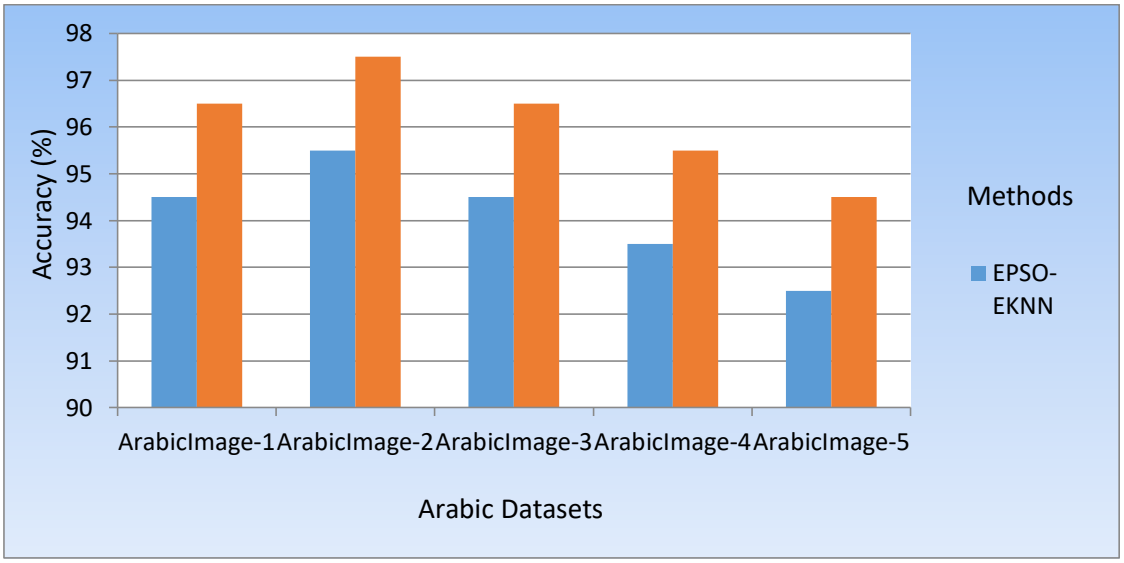

Figure 4: Graphical Accuracy Comparision

Precision

Precision refers to the ratio of the number of CR 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)"

Table 3
Numerical Precision Comparison

Arabic-Datasets	EPSON-EKEN	APCA-FKNN
ArabicImage-1	95	97
ArabicImage-2	96	98
ArabicImage-3	95	97
ArabicImage-4	94	96
ArabicImage-5	93	95

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 5 show the maximum precision of the CR was lower for AOCR by EPSO-EKNN and higher for AOCR by APCA-FKNN.

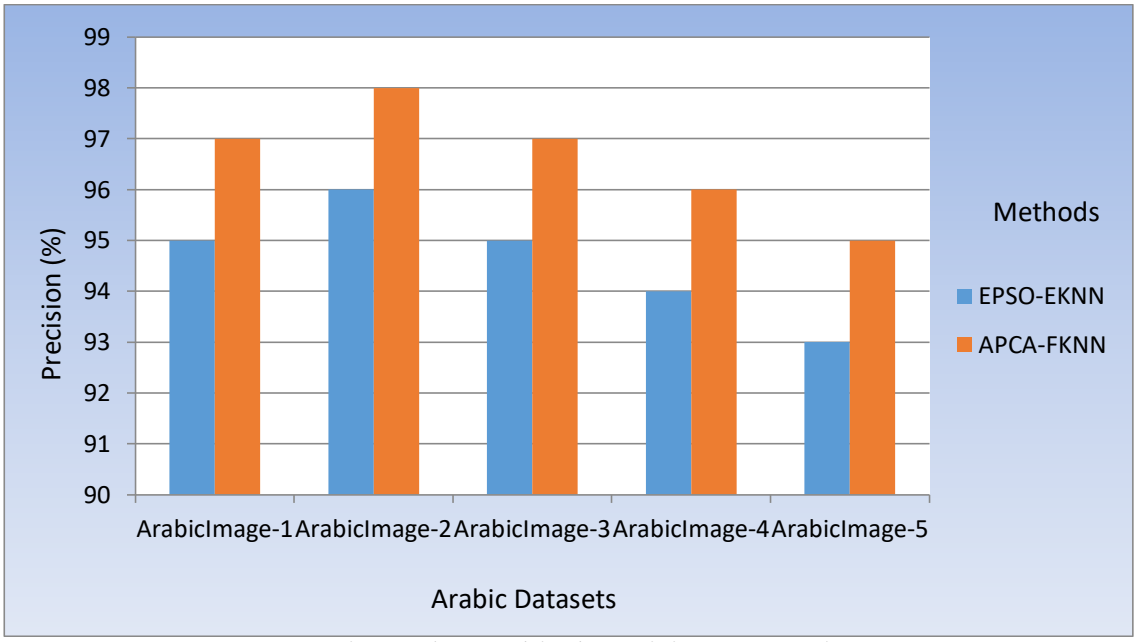

Figure 5: Graphical Precision Comparison

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)"

Table 4
Numerical Recall Comparison

Arabic-Datasets	EPSO-EKNN	APCA-FKNN
ArabicImage-1	96.5	97.5
ArabicImage-2	97.5	98.5
ArabicImage-3	96.5	97.5
ArabicImage-4	95.5	96.5
ArabicImage-5	94.5	95.5

This procedure is carried out on all images in the datasets. The findings are produced after evaluation upon CR utilizing actual reality information. Table 4 and Figure 6 show the maximum recall rate of the CR was lower without selecting optimal features for AOCR by EPSO-EKNN and higher for AOCR by APCA-FKNN.

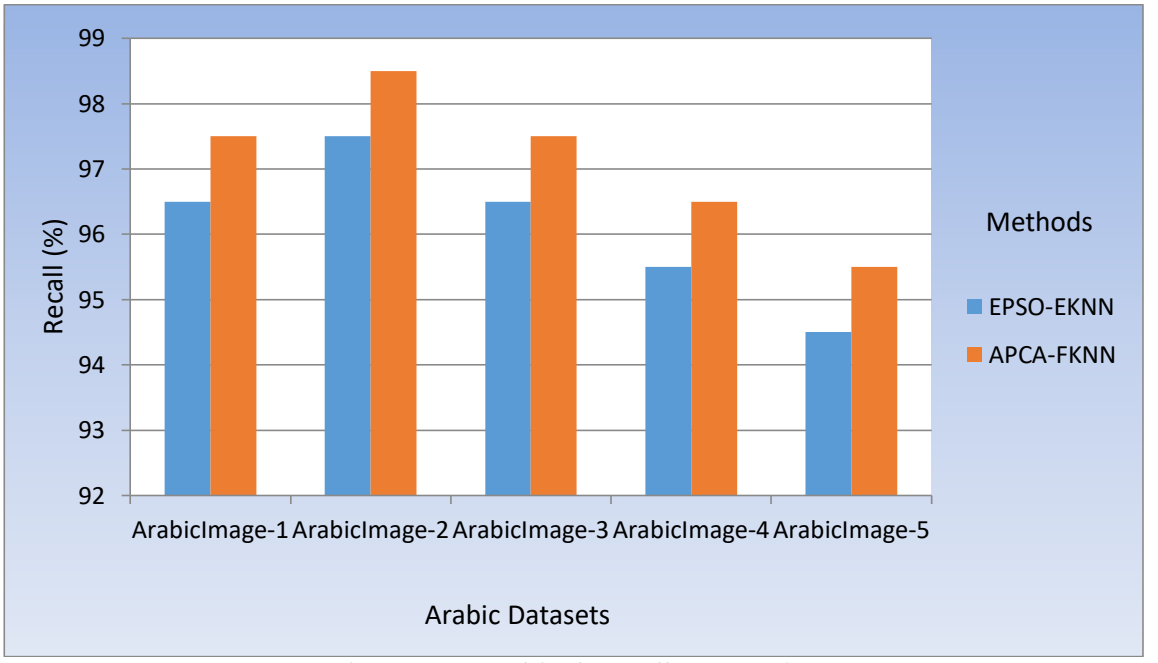

Figure 6: Graphical Recall Comparison

Conclusion

Research into the AOCR for textual information is a hard and open-ended attempt. Using the APCA for extracting features and the FKNN for classification, this article proposes a novel AOCR for individual characters. For its first iteration, it employed a given dataset image to train the model. The extraction of features procedure begins after the image has been denoised and the characters have been segmented. APCA is used in extracting the features with its capacity of using knowledge accumulated about an undiscovered searching space in an attempt to bias future searches towards potential subspaces. Also with the cursive form of Arabic-Characters in mind, the proposed model uses FKNN as a classification method. Fuzzification guarantees that vote from distinct samples corresponding to some more than a single class are counted with the membership function, which might be called proportional voting. The model's goal is to achieve the lowest identification error, the fastest runtime, and the smallest structure possible. In a limited period, the model can correctly identify 97.5 percent of various samples. In the presented classification approach, fuzzy classification is used to handle Arabic characters that overlap. To increase the efficiency of the classification, the FKNN method employs the semantic merging of both histograms and fuzzy's nearest neighbors to generate the membership function. Diacritics problems may be solved in the future with more powerful classification methods.

References

[1] S. Elsaid, H. Alharthi, R. Alrubaia, S. Abutile, R. Aljres, A. Alanazi, and A. Albrikan(2019) "Arabic real-time license plate recognition system", in Proc.1st Int. Conf. Comput. pp. 126-143.

[2] M. Awel, M. Ahmed, and A. Abidi,(2019) "Review on optical character recognition", Int. J. Comput. Appl., vol. 6, no. 5, pp. 3666-3669,

[3] F. Solamani and A. Mohamed,(2019) "Off-line optical character recognition system for Arabic handwritten text", J. Pure Appl. Sci., vol. 18, pp. 52-58,

[4] A. Lawgali,(2015) "A survey on Arabic character recognition", Int. J. Signal Process., Image Process. Pattern Recognit., vol. 8, no. 2, pp. 401-426

[5] I. A. Doush, F. AIKhateeb, and A. H. Gharibeh(2018) "Yarmouk Arabic OCR dataset", in Proc. 8th Int. Conf. Comput. Sci. Inf. Technol. (CSIT), pp. 150-154.

[6] Ashiquzzaman A, Tushar AK, Rahman A, Mohsin F (2019) An efficient recognition method for handwritten Arabic numerals using CNN with data augmentation and dropout. In: Balas VE, Sharma N, Chakrabarti A (eds) Data management, analytics, and innovation, advances in intelligent systems and computing. Springer, Singapore, pp 299–309

[7] R. Ahmad, S. Naz, M. Afzal, S. Rashid, M. Liwicki, and A. Dengel,(2020) "A deep learning-based Arabic script recognition system: Benchmark on KHAT", Int. Arab J. Inf. Technol., vol. 17, no. 3, pp. 299-305,

[8] Alani A (2017) Arabic handwritten digit recognition based on restricted Boltzmann machine and convolutional neural networks. Information 8(4):142.

[9] M. M. Mohamad, H. Hassan, D. Nasien, and H. Haron,(2015) "A review on feature extraction and feature selection for handwritten character recognition", Int. J. Adv. Comput. Sci. Appl., vol. 6, no. 2, pp. 204-213,

[10] Mars A, Antoniadis G (2016) Arabic online handwriting recognition using neural network. Int J Artif Intell Appl(IJAIA) 7(5)

[11] El-Sawy A, Loey M, Hazem E (2017) Arabic handwritten character recognition using convolutional neural network. WSEAS Trans Comput Res 5:11–19

[12] Maalej R, Kherallah M (2018) Convolutional neural network and best for offline Arabic handwriting recognition. In: 2018 International Arab conference on information technology (ACIT), Werdanye, Lebanon, 2018, pp 1–6. https://doi.org/10.1109/ACIT. 2018.8672667

[13] Latif G, Alghazo J, Alzubaidi L, Naseer MM, Alghazo Y (2018) Deep convolutional neural network for recognition of unified multi-language handwritten numerals. In: 2018 IEEE 2nd International

workshop on Arabic and derived script analysis and recognition (ASAR), pp 90–95

[14] F. Alotaibi, M. T. Abdullah, R. B. H. Abdullah, R. W. B. O. K. Rahmat, I. A. T. Hashem, and A. K. Sangaiah,(2018)``Optical character recognition for quranic image similarity matching,'' IEEE Access, vol. 6, pp. 554-562

[15] Dr E J Thomson Fredrik, V. M. Ashiq (2021). An OCR for Arabic Character Recognition with Ensemble Approach based Feature Selection for Enhanced KNN Classification. Design Engineering, 4728-4750. Retrieved from http://www.thedesignengineering.com/index.php/DE/article/view/5425.