# Literature Review

In recent years, the recognition of Arabic letters has become necessary in several applications, particularly in the context of this epidemic, since traditional education was mostly transferred to online learning, which may involve distance learning through the Internet for individuals who want to learn Arabic. In addition, it is very hard to identify the Arabic handwritten language, particularly in order to assess the accurate way of writing Arabic characters for postgraduate and undergraduate students, which requires scientific research (Shams et al., 2020).

The ability of a machine to recognize human handwritten input is known as automatic handwriting recognition. Handwriting can come from a variety of ways, including paper reports, photos; touch screens, and other electronic devices. Scanned handwriting is considered offline, whereas handwriting entered with a pen tip is considered online. The major challenges in handwriting letters recognition relate irregularities and variability of pattern, so the extraction of features is of prime significance. Identifying features manually may result in insufficient information being available to correctly estimate the character class. Even so, due to increased dimensionality, a large number of features usually lead to problems ( Altwaijry & Al-Turaiki, 2020).

Information retrieval systems and text mining, particularly search engines, require the recognition of characters and letters. There are many efforts at the present time to recognize, identify, verify and classify Arabic handwritten letters using machine learning (ML) methods (Shams et al., 2020). Many research studies in the fields of artificial intelligence (AI), computer vision (CV), and public relations (PR) considered handwritten character recognition (HCR) as an important area of research. These models generate strong conversions in simple machine learning (ML) applications when combined with artificial neural networks (ANNs) such as multi-layer perceptron (MLP) or long short-term memory (LSTM) network and a language model. Deep neural networks DNN, particularly convolutional neural networks (CNNs), are made up of many hidden layers with different architectures that are used to train a model which can correctly recognize words and reduce error rates significantly (Balaha et al.,2020).

1. **Arabic Language**

Arabic is the most widespread spoken Semitic language with an approximately 300 million speakers. It is currently ranks as the fifth most widely spoken language around the world. Arabic Handwriting recognition is still an open issue for a number of reasons. The cursive Arabic alphabet is used to write Arabic from right to left. The Arabic language alphabet has 28 letters, each of which has two to four different shapes depending on where it appears in a word. Furthermore, diacritical marks are used in Arabic writing to define short vowels and other sounds, such as “Alfatha”, “Al-dhamma”, and “Alkasra”. In addition, Arabic has several ligatures, composed of a combination of 2 or more letters, like: "alif-laam". ( Altwaijry & Al-Turaiki, 2020).

The most popular theory states that Arabic letters is derived from Nabatean script. However, Diacritical marks were not available in the Nabataean alphabet.Secondary components have been developed by Arab philologists to distinguish letters with similar shapes but different phonetics. These are known as diacritical marks, and they may be points (one, two, or three), vocalization signs, or other signs like Alhamza, Alshadda or Almadda, and so on. ( Lamghari & Raghay, 2021). Diacritical points only appear on or under the letter. The highest number of points that may have a character is 3 points above the letter or 2 points below such as:

ش ، ي ، ق ،ث ، ت، ج، خ ،ش ،ض ،ظ

Vowels in Arabic are not letters but diacritical marks linked with the letters to which they relate. There are 3 kinds of vowels:  long vowels such as ( ا، و، ي), short vowels ( َ، ُ، ِ) and doubles ( ٌ، ً، ٍ). Vowels are very essential in Arabic, and the absence of vowels may cause misunderstanding the sentence since the word's meaning can change if the proper vowel is missing ( Lamghari & Raghay, 2021). For example

كَتَبَ كٌتُب

Book wrote

1. **Convolutional Neural Networks**

Convolutional neural networks (CNN) are a type of deep learning neural networks. CNN network is usually used in videos and image recognition, classification of image, natural language processing (NLP), and also in medical imaging. Also, It is called space invariant or shift invariant artificial neural networks (Morsy et al.,2020).

A CNN is a feed-forward multi-layer neural network that extracts features and characteristics from input data (sounds or images). The back-propagation algorithm was used to train the CNN. CNNs can learn nonlinear mappings from a large number of data sets with high-dimensional complex inputs (sounds or images).The benefit of CNN is that it automatically extracts the key features which are invariant and a certain extent to shapes and shifts the distortions of the input letters. A further main advantage of CNN is the utilize of shared weight in convolutionary layers, that means that the same filter is being utilized for every layer input. Share weight decreases the parameters number and increases performance (El-Sawy et al.,2017).

1. **Arabic Character Recognition System**

Arabic handwritten character recognition systems will take one of two directions. The method of classifying a character during writing is known as online character recognition. This method needed a special kinds of pen and tablet such as a mobile phone or a personal digital assistant (PDA). Offline character recognition systems, on the other hand, work with pre-scanned photos and papers. There are three different types of offline character recognition. The first is the identification of written letters with similar shapes, sizes, and fonts, which makes it the most straightforward of the offline recognition systems. The second is the identification of handwritten letters in which each user's style varies. The third form of character recognition is typeset letter recognition. Recognition of handwritten text can be done in two different ways: without segmentation methods and with segmentation methods. Due to the ligatures and cursive form of Arabic scripts, it is preferable to identify the whole word in order to prevent a significant increase in error rates (Balaha et al.,2020).

The Arabic handwritten recognition process can be classified into different phases as follow:

* **Image retrieval (acquisition) phase.** The aim of this stage is to retrieve and transform information (images) from various sources into a digitalized context.
* **Preprocessing phase.** The aim of this phase is to enhance the textual image readability and eliminate unnecessary features from it.
* **Data augmentation.** It is a method that permits us to raise the variety and size of data available for training sets without any new data collected.
* **Segmentation phase.** Texts are segmented into smaller units during the segmentation process. Words, Lines, and characters can all be used.
* **Feature extraction phase.** Features are extracted and captured during this step. This is regarded as one of the most risky stage of OCR.
* **Classification phase.** The process of recognizing and identifying an entity by comparing its characteristics to one of the classes is known as classification.
* **Post-processing phase.** This phase is in responsible of producing the best solution and is often applied as a collection of methods and tools that primarily depend on letter lexicons, frequency, and other data.

1. **Related works**

Recent studies on Arabic character recognition systems are reviewed in this section. In comparison to Latin and Chinese languages, little research has been done in the domain of Arabic handwritten character recognition. The literature review demonstrate that there are two types of studies in the field of Arabic recognition system: printed characters recognition systems and handwritten characters recognition systems. In order to design a recognition system, a characters database is required to establish the recognition process. Due to the lack of a publicly available database reference for Arabic character, many  researchers have design their systems using a set of data they have collected themselves.

* 1. **Databases**

Altwaijry & Al-Turaiki (2020) introduced "Hijja" dataset, a new dataset of Arabic letters written by kids aged between 7–12 years. This dataset consist of 47,434 letters that written by 591 children. Also, they suggested an automatic handwriting recognition system on the basis of convolutionary neural networks (CNN).  They trained the system on " Hijja" dataset as well as the Arabic Handwritten Dataset (AHCD). The findings demonstrate the effectiveness of the suggested system is encouraging, with the accuracy of 88% and 97% for the Hijja dataset and the AHCD dataset, respectively.

AbdElNafea and Heshmat (2020) proposed "Aswan online Arabic handwritten" (AOLAH) databases, which are new databases for Arabic handwriting recognition.The first database that has been suggested includes 2,520 files that describe 28 Arabic letters written by 90 different authors. Those who have contributed are engineering students from Aswan University. The other database is generated by extracting only the strokes from the primary database.There are 1,530 files in  second database for 17 strokes. The suggested data collection algorithm is special in that it treats each stroke in the character separately. Stroke capturing involves gathering data points along the path of an input tool (stylus pen or mouse) as certain characters are being drawn.

Lamghari & Raghay (2021) introduced a novel database for handwritten Arabic diacritics (DBAHD).  It is intended to serve Arabic handwriting recognition systems based on machine learning and segmentation.This database includes 500 types of diacritics that can be utilized to evaluate the efficiency of many techniques and methods for the recognition of handwritten Arabic letters that involve diacritical marks, in specific those that identify the word after segmentation.

**4.2 Arabic letters recognition systems**

Shams et al.(2020) addressed the issue of recognition of Arabic hand-written letters by evaluating the similarity between input scripts and prestored templates by utilizing fully-connected deep convolution neural networks (DCNN) and drop-out support vector machines (SVM). In addition, they determined the rate of correct classification (CRR) and the rate of error classification (ECR).The experimental outcomes of this study confirm the efficiency of the suggested algorithm to identify recognize and check the input of handwritten Arabic letters. In addition, the suggested model defines similar Arabic characters using a clustering model based on the K-means clustering methods to address the multi-stroke problem in Arabic letters. The comparative assessment is indicated and the accuracy of the system was 95.07 percent for CRR with 4.93 percent for ECR compared to the state of the art.

Tang et al.,(2021) introduced a method to recognize handwritten Arabic numerals from zero to nine by utilizing a Probabilistic Neural Network (PNN) method. This method, which includes two image pre-processing and recognition processes, can identify input written by hand and imported externally Arabic numerals in actual time. Normalization and expansion is used in image pre-processing to expand the picture's characteristics for easier identification. This method has great performance and reusability, so it can be used as a simple testing platform for further research into handwritten numeral Arabic recognition.

Can & Kabaday (2020) applied an automated Arabic numeral spotting scheme to a sample of the Ottoman Empire's first population registers, which were collected in the mid-nineteenth century. They relied on the property that the numerals in community registers are displayed in colour. They proposed a CNN-based numeral spotting system after adding a red color mask.They also mapped the local Arabic handwritten digit set of data from the spotted numbers by choosing unidigit ones and evaluated the Deep Transfer Learning (DTL) approach from big open Handwritten arabic digit sets of data for recognition of digit. The CNN model was trained on the local set of data and evaluated on a different testing data that outperforms Deep Transfer Learning (DTL) methods with a digit recognition accuracy of 80 per cent.

Balaha et al.,(2020) introduced a deep learning (DL) system consists of two convolution neural network (CNN) models (named HMB1 and HMB2) with a big and complex set of data for Arabic handwritten character recognition (AHCR)  problems. They concluded from the findings that data augmentation aided in improving testing accuracy and reducing overfitting.HMB1 had higher accuracies than HMB2 because the former's architecture is more complex and contains more parameters, but it takes longer to learn. Using the CMATER dataset with and without data augmentation, the two mentioned architectures accomplished more than 99.2 percent and 100 percent, respectively.

Rahal et al. (2021) presented  a novel deep learning-based method for recognizing Arabic text in pictures. They suggested a new hybrid network that combines a Bag-of-Feature (BoF) architecture for extracting features and Hidden Markov Models (HMMs) to recognition of sequence. The suggested model, named BoF-deep SAE-HMM, is evaluated on 4 data collections. The experiments carried out demonstrated a significant robustness of the recognition system, whatever the database used.

Al-wajih & Ghazali (2020) used sliding windows to improve classification accuracy and applied random forest (RF) and vector support (SVM) classification methods for the identification of Arabic digit pictures. In order to investigate their efficiency with and without sliding windows, 4 distinct extraction techniques, including "Gray-Level Co-occurrence Matrix (GLCM)", "Moment-based", "Mean-based", and "Edge Direction Histogram (EDH)", have been proposed. The observed accuracy demonstrates the importance utilizing sliding windows for digit classification. The rate of recognition obtained by utilizing modified version of the "AHDBase" dataset are 98 per cent when Moment-based and Mean based are utilized with the RF classifier, 98.33 per cent and 99.13 per cent respectively when GLCM and EDH are utilized with the linear kernel SVM.

Morsy (2020) proposed a new Arabic handwriting recognition (AHDR) method based on CNN and an entity property called entropy. The entropy is an objective function for evaluating the performance accuracy as a function of the network's weights and biases.The results indicated that the proposed technique provided a low error rate and high accuracy in comparison to other techniques.

Noubigh et al.,(2020) proposed a model using the CNN-BLSTM-CTC architecture to perform a text recognition task for Arabic scripts using a the technique of transfer learning . The proposed solution consists of two steps that include extraction features and modeling sequences. The experiments have shown a high performance of the proposed scheme and Strong results are achieved by transferring learning technique.

# References

[1] Shams, M., Elsonbaty, A., & ElSawy, W. (2020). Arabic Handwritten Character Recognition based on Convolution Neural Networks and Support Vector Machine. *arXiv preprint arXiv:2009.13450*.

[2] Altwaijry, N., & Al-Turaiki, I. (2020). Arabic handwriting recognition system using convolutional neural network. *Neural Computing and Applications*, 1-13.

[3] Tang, Y., Zhang, S., & Niu, L. (2021). Handwritten Arabic numerals recognition system using probabilistic neural networks. In *Journal of Physics: Conference Series* (Vol. 1738, No. 1, p. 012082). IOP Publishing.

[4] Can, Y. S., & Kabadayı, M. E. (2020). Automatic CNN-Based Arabic Numeral Spotting and Handwritten Digit Recognition by Using Deep Transfer Learning in Ottoman Population Registers. *Applied Sciences*, *10*(16), 5430.

[5] Balaha, H. M., Ali, H. A., Saraya, M., & Badawy, M. (2020). A new Arabic handwritten character recognition deep learning system (AHCR-DLS). *Neural Computing and Applications*, 1-43.

[6] Rahal, N., Tounsi, M., Hussain, A., & Alimi, A. M. (2021). Deep Sparse Auto-Encoder Features Learning for Arabic Text Recognition. *IEEE Access*, *9*, 18569-18584.

[7] Balaha, H. M., Ali, H. A., & Badawy, M. (2020). Automatic recognition of handwritten Arabic characters: a comprehensive review. Neural Computing and Applications, 1-24.

[8] Al-wajih, E., & Ghazali, R. (2020). Improving the accuracy for offline Arabic digit recognition using sliding window approach. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, *44*(4), 1633-1644.

[9] AbdElNafea, M., & Heshmat, S. (2020). Novel Databases for Arabic Online Handwriting Recognition System. In *2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE)* (pp. 263-267). IEEE.

[10] Morsy, H. A. (2020). Optimization of Arabic Handwritten digits recognition using CNN.

International Journal of Scientific & Engineering Research ,11( 11), 372 .

[11] Noubigh, Z., Mezghani, A., & Kherallah, M. (2020). Transfer Learning to improve Arabic handwriting text Recognition. In *2020 21st International Arab Conference on Information Technology (ACIT)* (pp. 1-6). IEEE.

[12] Lamghari, N., & Raghay, S. (2021). Recognition of Arabic Handwritten Diacritics using the new database DBAHD. In *Journal of Physics: Conference Series* (Vol. 1743, No. 1, p. 012023). IOP Publishing.

[13] El-Sawy, A., Loey, M., & El-Bakry, H. (2017). Arabic handwritten characters recognition using convolutional neural network. *WSEAS Transactions on Computer Research*, *5*, 11-19.