

## Article

# Machine Learning Approach for Arabic Handwritten Recognition

A. M. Mutawa <sup>1,2,\*</sup> , Mohammad Y. Allaho <sup>1</sup>  and Monirah Al-Hajeri <sup>1</sup>

<sup>1</sup> Department of Computer Engineering, College of Engineering and Petroleum, Kuwait University, Safat 13060, Kuwait; m.allaho@ku.edu.kw (M.Y.A.)

<sup>2</sup> Computer Sciences Department, University of Hamburg, 22527 Hamburg, Germany

\* Correspondence: dr.mutawa@ku.edu.kw

**Abstract:** Text recognition is an important area of the pattern recognition field. Natural language processing (NLP) and pattern recognition have been utilized efficiently in script recognition. Much research has been conducted on handwritten script recognition. However, the research on the Arabic language for handwritten text recognition received little attention compared with other languages. Therefore, it is crucial to develop a new model that can recognize Arabic handwritten text. Most of the existing models used to acknowledge Arabic text are based on traditional machine learning techniques. Therefore, we implemented a new model using deep machine learning techniques by integrating two deep neural networks. In the new model, the architecture of the Residual Network (ResNet) model is used to extract features from raw images. Then, the Bidirectional Long Short-Term Memory (BiLSTM) and connectionist temporal classification (CTC) are used for sequence modeling. Our system improved the recognition rate of Arabic handwritten text compared to other models of a similar type with a character error rate of 13.2% and word error rate of 27.31%. In conclusion, the domain of Arabic handwritten recognition is advancing swiftly with the use of sophisticated deep learning methods.

**Keywords:** machine learning; handwritten recognition systems; Arabic handwriting; BiLSTM; ResNet; natural language processing



**Citation:** Mutawa, A.M.; Allaho, M.Y.; Al-Hajeri, M. Machine Learning Approach for Arabic Handwritten Recognition. *Appl. Sci.* **2024**, *14*, 9020. <https://doi.org/10.3390/app14199020>

Academic Editor: Douglas O'Shaughnessy

Received: 30 August 2024

Revised: 19 September 2024

Accepted: 1 October 2024

Published: 6 October 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Optical character recognition (OCR) is a field of research in pattern recognition (PR). The goal of an OCR system is to automatically read text from a scanned paper and convert it to a digital format that can be readable and editable using electronic applications [1].

Currently, the technologies are spread worldwide, and almost all the essential processes are being performed electronically. Also, the Arabic language is the official language of 23 countries and is spoken by more than 400 million people worldwide [2]. This raises the need for an efficient Arabic text recognizer, which can be helpful in many institutions, such as educational, governmental, and economic organizations. For example, it is essential to convert old and/or new documents with handwritten scripts into digital Arabic text in some institutions. The OCR system helps professionally complete office tasks while saving time and effort. Moreover, recognizing Arabic handwritten text is helpful in the procedure of the automatic reading of bank checks [3].

It is challenging to recognize the human writing of Arabic text because of the cursive nature of Arabic handwriting scripts. The different shapes of Arabic letters depend on their location in the word and the special marks used in some Arabic letters, such as 'Hamza' and 'Maada'. Also, a lot of Arabic letters have the same shape but are differentiated by dots, which can be one, two, or three dots placed either above or below the character [4].

The Arabic script is written from right to left, so it is essential to recognize the words in the same direction. Due to the challenges of recognizing Arabic text and the various characteristics of Arabic writing from other writings, it is difficult to apply the techniques

developed for identifying other languages in Arabic scripts. Therefore, we implemented a new model that will recognize offline Arabic handwritten text.

Arabic text recognition systems can be either based on segmenting the word (analytical approach) or without segmentation. Most of the current systems are segmentation-based systems, which require segmenting the word into different characters [5]. Then, after the segmentation process, each character is recognized. However, due to the cursive nature of the Arabic handwritten text, it is challenging to segment words into characters [6]. On the other hand, segmentation-free models (holistic approach) recognize words as a whole-word images without any segmentation processes. The holistic approach is preferred for data with small vocabulary sizes, such as the recognition of bank checks, while the analytical method is ideal for recognition systems that consist of large vocabularies [7,8].

Traditional approaches to developing Arabic handwritten recognition systems are based on shallow learning techniques. Using these techniques makes it difficult to deal with the challenges of recognizing Arabic handwritten words. This is because they only extract the sample features of the word image. However, deep machine learning approaches have better performance in many systems since they can extract more complex features from the word image [9,10]. Thus, using machine learning approaches is helpful to handle the challenges of recognizing Arabic handwritten words [11,12].

The identification of Arabic characters poses persistent challenges due to several factors, and continuous research is being conducted to enhance the performance of current systems. Several methodologies are constrained to proprietary datasets or the identification of individual words or paragraphs, which complicates the evaluation of their effectiveness on authentic Arabic literature [13].

Within the field of deep learning, multiple architectures have become quite effective instruments for different purposes, including image and text recognition. ResNet is a novel convolutional neural network (CNN) architecture that solved the vanishing gradient issue, therefore enabling the training of very deep networks and hence the idea of residual learning [14]. The BiLSTM networks acquire contextual information and long-term associations from sequences by processing data both forward and backwards. These networks are a type of recurrent neural network. Text recognition and natural language processing are two examples of tasks that benefit greatly from this feature of BiLSTM [15,16].

When dealing with sequence-to-sequence problems, if the alignment of the input and output sequences is unknown, the CTC technique is a useful complement to these structures [17]. The ability to predict sequences of varying lengths without utilizing pre-segmented input is a common use case for CTC, which finds widespread use in voice and handwriting recognition. The combination of CNN's feature extraction capabilities, RNNs' sequence modeling capabilities, and CTC's adaptive sequence alignment capabilities provides a solid basis for tackling difficult recognition tasks [18].

This work implements a segmentation-based model using deep machine learning techniques to have a high accuracy rate. The system is evaluated using King Fahd University of Petroleum and Minerals (KFUPM) Handwritten Arabic Text database (KHATT) [19] and the Arabic Handwritten Text Images Database written by Multiple Writers (AHTID/MW) datasets [20]. These are challenging datasets that contain text line images. These datasets cover different writing styles of other writers; the system will recognize Arabic handwritten text and words from a text line image.

The contributions of the study are as follows:

- We implemented the ResNet model to extract the features of every individual character in the textual image. The BiLSTM with CTC model was employed for the purpose of sequence modeling. Ultimately, a language model (LM) was employed during the post-processing phase to improve the forecasted outcome derived from the classification phase.
- We performed testing of the model on two distinct datasets, the KHATT and AHTID/MW datasets. The utilization of several datasets underscores the model's capacity to extrapolate across diverse manifestations of Arabic handwriting.

The subsequent sections of the article are classified as follows: Section 2 presents the literature review of OCR systems, while Section 3 provides an outline of the study approach. The findings are articulated and examined in Section 4. Section 5 concludes the study, highlighting specific limitations and prospective areas for further research.

## 2. Literature Review

Many optical character recognition systems have been designed to recognize Arabic handwritten characters and words. CNNs have been widely employed in handwritten recognition due to their capacity to autonomously acquire hierarchical features from unprocessed pixel input [21,22]. A combination of two classifiers, which are CNN and Support Vector Machine (SVM) with a dropout regularization technique, is used to recognize offline Arabic handwriting characters [23]. The authors use SVM to adjust the trainable classifier of CNN. The dropout is applied before supplying the output into an SVM classifier. The authors tested their design on a character dataset, a handwritten Arabic characters database (HACDB), and a word dataset that is IFN/ENIT. The experimental results showed that the HACDB dataset had a 5.83% error classification rate. The IFN/ENIT dataset had a 7.05% error classification rate.

An Arabic handwriting recognition system was proposed based on multiple BiLSTM-CTC combinations [24]. In this study, two different extraction techniques were used. The first method is segment-based features. The second is Distribution-Concavity (DC)-based features trained on different levels of the BiLSTM-CTC combination. The combination levels were low-level fusion, mid-level combination methods, and high-level fusion. The experiments were performed on the KHATT dataset. The results showed that the high-level fusion had a better recognition rate than the other combination levels, with a 29.13% word error rate (WER) and a 16.27% character error rate (CER).

BenZeghiba, Louradour, and Kermorvant used a hybrid Hidden Markov Model (HMM) and Artificial Neural Network (ANN) framework to recognize Arabic handwritten text [25]. The type of ANN used in their system is Multi-Dimensional Long Short-Term Memory Networks (MDLSTMs). The hybrid model extracts the pixel values of text line images by scanning the text line images in four directions. A CTC is used during the training process. The Viterbi algorithm [26], a decoding strategy, is used to generate the best hypothesis of a character sequence. They added a hybrid language model that consists of words and Part-of-Arabic Words (PAWs). The KHATT dataset was used to evaluate the system, and the result was a 33% WER.

A recognition system for Arabic handwritten text was proposed by Stahlberg and Vogel [27]. A sliding window is used to extract features from text-line images. The window's width is 3 pixels with an overlap of 2 pixels. The parts are extracted using two different strategies. The first strategy is pixel-based features extracted from raw grayscale pixel values. The second strategy is segment-based features, consisting of centroid and height features. Kaldi toolkit, which is used in speech recognition systems and is based on deep neural networks, is used for classification [28]. The best word error rate was obtained from pixel-based features with a 30.5% WER on the KHATT corpus.

Wigington et al. introduced two data augmentation and normalization methods: a novel profile normalization strategy for both word and line images and an augmentation of existing text images using random perturbations on a regular grid [29]. These techniques were used with a CNN-LSTM architecture to enhance handwriting text recognition. Contemporary youngsters frequently utilize technology, and their distinctive characteristics in handwriting differ from those of adults. Therefore, the study by Altwaijry et al. [30] has been trained on children's handwriting.

The work by Khayati et al. [31] emphasizes the development and efficiency of several CNN architectures in tackling the distinct difficulties presented by Arabic script, such as writing with cursive and the existence of diacritical markings. Lamia et al. [11] developed a CNN-graph theory method for Arabic handwritten character segmentation. They address the difficulty of segmenting linked and overlapping cursive Arabic letters, a major obstacle

to effective character identification. In a study by AlShehri [32], a deep neural network for Arabic handwritten recognition (DeepAHR) improves feature extraction and recognition with a complex neural network design. The model excels at Arabic script character segmentation and contextual changes. DeepAHR outperformed prior models in accuracy and processing speed.

In another study by Alghyaline [33], the Arabic handwritten recognition was implemented using different CNN pretrained models such as Visual Geometry Group (VGG), ResNet, and Inception on three different datasets: Hijja, the Arabic Handwritten Character Dataset (AHCD), and the AlexU Isolated Alphabet (AIA9K). Each dataset achieved accuracies of 93.05%, 98.30%, and 96.88% on the VGG model. The transformer transducer and the typical transformer design that makes use of cross-attention were the two end-to-end architectures that were explored in the work by Momeni and BabaAli [34]. They employed the KHATT dataset and obtained a CER of 18.45%. Table 1 describes the recent works on OCR.

**Table 1.** Background studies based on Arabic handwritten recognition with deep learning.

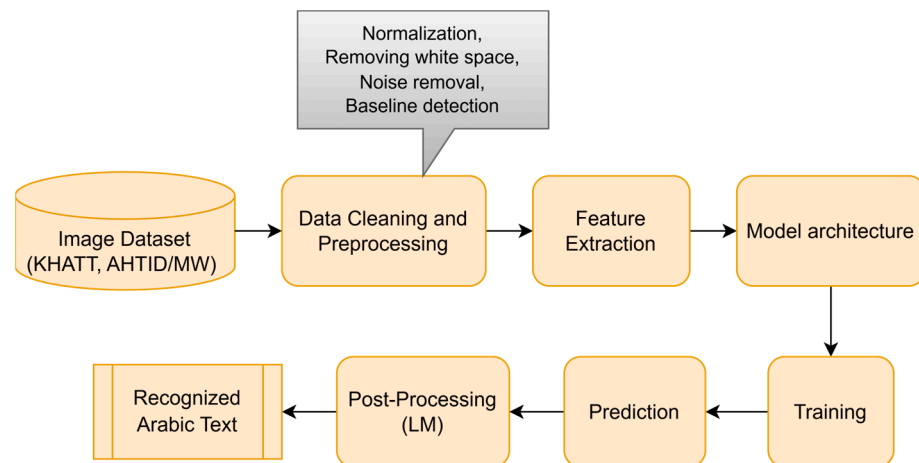
Reference, Year	Dataset	Method	Result	Limitation
[25], 2015	KHATT Maurdor	MDLSTM	31.3% WER 33.2% WER	Requires more memory and computational resources
[27], 2015	HACDB IFN/ENIT	Kaldi's network model	30.5% WER 11.5% WER	No focus on the LM
[23], 2016	HACDB IFN/ENIT	CNN-SVM	5.83% ECR 7.05% ECR	Overfitting, computational complexity of SVM
[24], 2018	KHATT	BiLSTM	29.13% WER 16.27% CER	Limited vocabulary
[30], 2021	Hijja AHCD	Segmentation, CNN	88% Accuracy 97% Accuracy	Limited to characters written by children
[11], 2024	IESK-ArDB	SHG, CNN	96.97% Accuracy	Computational complexity due to the number of edges in the graph
[12], 2024	IFN/ENIT	CNN-BiLSTM	4.012% WER	Model's generalizability
[32], 2024	Hijja AHCD	Segmentation, CNN	88.24% Accuracy 98.66% Accuracy	Limited to characters written by children
[33], 2024	Hijja AHCD AIA9K	VGG	96.88% Accuracy 98.30% Accuracy 93.05% Accuracy	Evaluation is limited to recognizing characters alone
[34], 2024	KHATT	Transformer with attention	18.45% CER	Model's generalizability

ECR: Error classification rate, IESK-ArDB: Institute for Electronics, Signal Processing and Communications Arabic database, SHG: segmentation hypothesis graph.

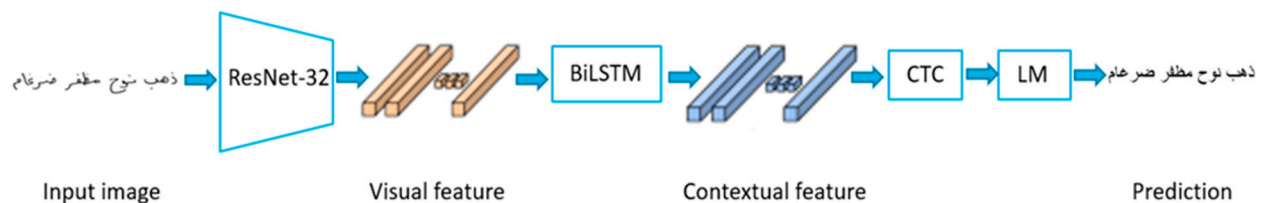
### 3. Materials and Methods

The Arabic handwritten text recognition system should have multiple stages to convert a handwritten text image into a digital format. This system consists of four consecutive processing stages: preprocessing, feature extraction, classification, and post-processing, as shown in Figure 1. The output of each process is used as an input to the process that follows.

First, preprocessing techniques are applied to the scanned image to improve the readability of the text. After that, the ResNet model is used to extract the features of each character in the text image. These features are the input for the classification stage. The BiLSTM-CTC network converts the visual features into contextual features. It predicts the sequence of characters with the help of the predefined classes in the database. Finally, an LM is used in the post-processing stage to enhance the predicted result from the classification stage. Figure 2 shows the workflow of our model's architecture. Each stage will be discussed in detail in the following subsections.



**Figure 1.** Block diagram of the proposed study.



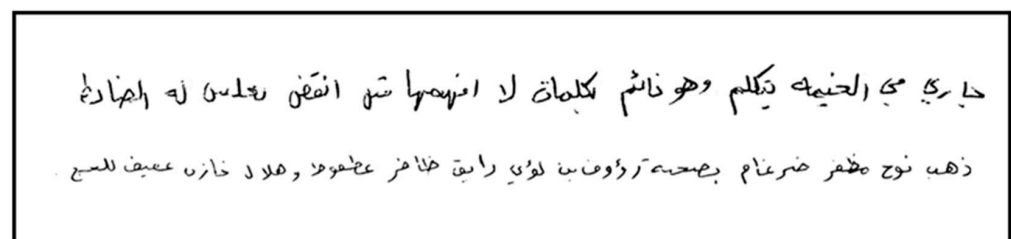
**Figure 2.** Our model's architecture.

### 3.1. Datasets Description

The number of available Arabic handwritten text databases is limited. In our model, we used two different datasets, the KHATT and AHTID/MW datasets, to train and test our model. These datasets contain all the Arabic characters written in different writing styles by different writers.

#### 3.1.1. KHATT

The KHATT is one of the challenging Arabic handwritten text databases published by KFUPM [19]. The KHATT is an offline Arabic handwritten text database consisting of text line images extracted from handwritten forms filled out by 1000 different writers. The writers are from different regions, gender, age, left/right-handedness, and educational background. The database consists of 300 Dots Per Inch (DPI) grayscale images of 2000 unique text paragraphs (randomly selected) and 2000 fixed text paragraphs (similar text). The database also contains 300 DPI binary text line images extracted from the paragraphs [35]. Figure 3 shows some samples from the KHATT dataset.



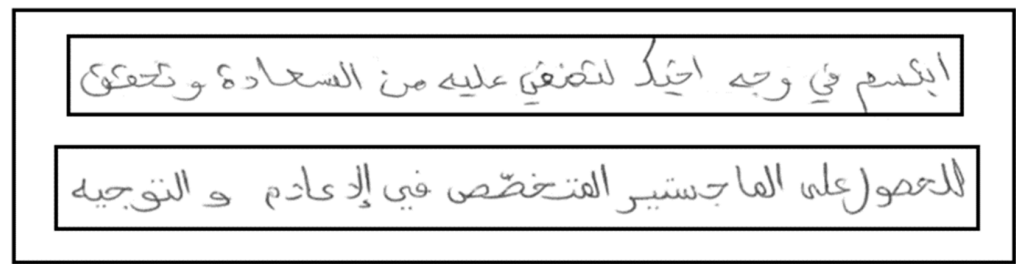
**Figure 3.** Samples from the KHATT dataset.

#### 3.1.2. AHTID/MW Database

The AHTID/MW dataset developed by [20] includes 3710 handwritten Arabic text lines and 22,896 words written by 53 native Arabic writers of different ages and educational



levels. The dataset consists of grayscale text line images with 300 DPI resolution. The dataset contains a variety of Arabic handwriting writing styles, as shown in Figure 4.



**Figure 4.** Samples from the AHTID/MW dataset.

With varied samples in terms of writing styles and significant data for training and assessment, both the KHATT and AHTID/MW datasets provide priceless resources for Arabic OCR research. Handling the cursive character of Arabic script, the existence of diacritics, and the variation in individual writing styles are the main difficulties related to these datasets. Confronting these problems is essential for creating efficient OCR systems that achieve accurate results across various handwriting samples. Table 2 shows the statistics of the dataset used for the study.

**Table 2.** Statistics of the datasets used in the study.

Dataset	Number of Lines	Number of Words	Number of Writers
AHTID/MW	3710	22,896	53
KHATT	13,357	185,272	1000

### 3.2. Preprocessing

The preprocessing step for scanned text images is very critical. It improves accuracy in handwritten text recognition systems. First, image binarization is applied to convert grayscale images. Each image uses values ranging from 0 to 256 for each pixel, which are converted to represent a black and white image represented by 1 and 0, respectively.

Arabic text line image datasets have high skew and extra white spaces. We removed the extra white regions by scanning the image from top to bottom to locate the first black pixel at the top and the position of the lowest black pixel at the bottom. After detecting the highest and lowest points of the black pixels, the text line images are cropped [36]. The exact process is repeated from the left to the right side. Figure 5 shows a sample image from the KHATT dataset after removing the white spaces.

Handwritten text is freestyle handwriting; therefore, noises or unwanted text such as random lines and dots may exist in the text image. Some images in the dataset contain a straight horizontal line on top of the text. In this work, we removed the horizontal line by computing an image difference in the horizontal direction, and from the image difference, we can find the horizontal line in the image by searching the continuous difference value. Doing so, we can filter and obtain the horizontal line by setting a threshold value to determine whether a horizontal area can be considered a line. Suppose the length of the horizontal line is greater than the image width size divided by 10. In that case, the filtered horizontal line is removed from the image. Figure 6 shows a sample image from the KHATT dataset where the upper horizontal line is removed from the text line image.

Noise filtering techniques are applied to remove noises from the text line images. Max and Min filters, also known as erosion (minimum) and dilation (maximum) filters, are used in the preprocessing stage to remove noises from the text images efficiently [37]. These filters are morphological transformation filters that define the neighborhood around each pixel. Erosion and dilation are two basic morphological operators [38].

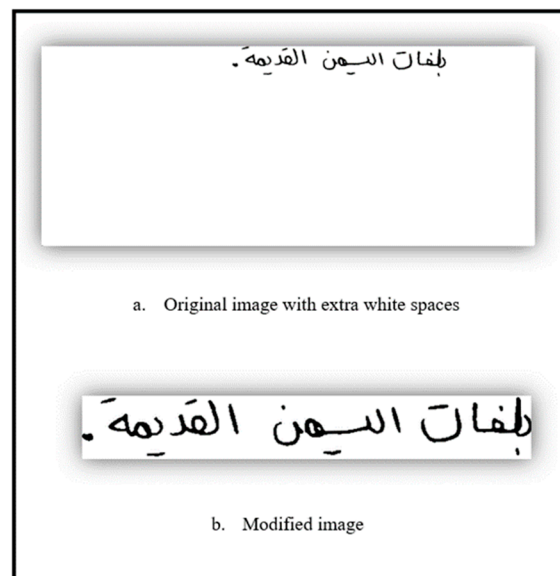


Figure 5. Sample image from the KHATT dataset after removing the white spaces.

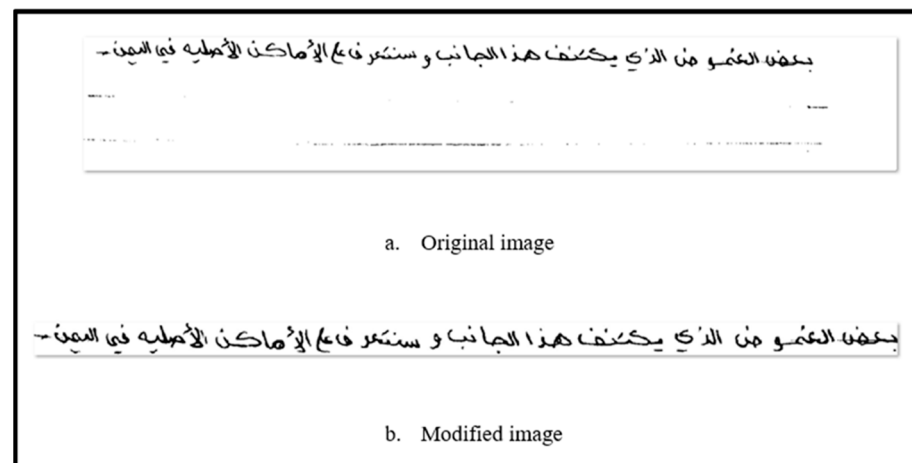


Figure 6. Sample image from the KHATT dataset after removing the upper line.

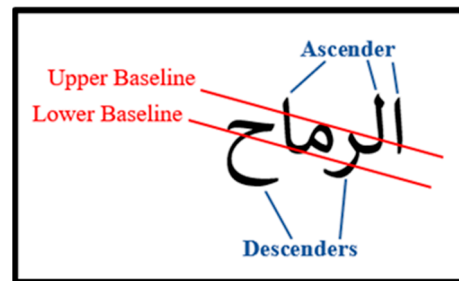
First, erosion is applied to the text image to erode the foreground object. It makes it smaller, that is, it removes small pixels (noises) near the boundaries of the foreground object (characters). Then, the dilation process is used to increase the size of foreground pixels (characters). We used erosion followed by dilation because erosion removes the noises in the text image, but it also shrinks the characters. Therefore, after the noises are removed from the text image, we dilate it. The dilation process enhances the distinctness of the characters and helps in joining broken parts of the text image together. An example of the Max and Min filter applied to a sample image from the KHATT dataset is shown in Figure 7.

Image normalization is performed, which helps reduce the text image's skewness and facilitates characters' visual learning features. Arabic text contains two baselines: the upper and lower baselines, as shown in Figure 8. The two baselines identify the core zone, the upper region with the ascenders, and the lower region includes the descenders. The core zone typically has a significant fraction of foreground pixels.

We used a method proposed by Stahlberg and Vogel [39] for baseline estimation by finding stripes in an image with a dense foreground. First, we detect the baseline for the whole image and rotate the image so that the detected baseline is horizontal. After that, we split the image vertically into smaller slices. Then, we detect the baseline for each slice of the image separately and rotate the sliced images such that the baseline becomes horizontal. Finally, we concatenate all the sliced images with a straight horizontal baseline to a single image.



**Figure 7.** Sample image from KHATT dataset after applying the Max and Min filter.



**Figure 8.** Upper and lower baselines in Arabic text.

### 3.3. Feature Extraction

Feature extraction is the second phase after the data have been preprocessed. Features are the main point around which the whole system is built. They are the target for all the previous stages and the input to the classification phase. Different feature extraction methods are applied in Arabic text recognition systems. Some approaches used hand-crafted feature extraction techniques using statistical [40] or structural [41] features. Other approaches computed both statistical and structural features [42–44].

Recently, a new trend has shifted from handcrafted feature extraction methods towards machine learning techniques for feature extraction and text recognition systems. Deep networks, one of the most advanced machine learning techniques, simulate human brain activity and automatically extract features from text images. However, deep models for Arabic handwritten text recognition systems are rare compared to other languages due to their complexity and cursive writing style [45]. A convolutional neural network (CNN) is an artificial neural network used in the pattern recognition field for image processing and recognition. CNN's have proven their effectiveness in understanding image content, providing state-of-the-art image recognition and detection [46]. Therefore, since CNN's have shown their ability to learn interpretable and powerful features from an image [47], we adopted a CNN architecture in our model to extract features from the text image.

In our system, we used a ResNet-based model [48], which is a robust CNN architecture, to extract features from the text images. The CNN state-of-the-art architecture goes deeper each year since Krizhevsky et al. [49] presented AlexNet 2012. While AlexNet consisted of only five convolutional layers, the VGG (Visual Geometry Group) network had 16–19 convolutional layers [50]. Googlenet consisted of 22 deep convolutional layers [51].

However, enabling the model to learn better and more features by increasing network depth is not as simple as stacking more layers together. Deep networks are hard to train because of the vanishing/exploding gradient problem, where, as the gradient is



backpropagated to earlier layers in the network, frequent multiplication might make the gradient infinitely small (i.e., vanish or explode) [52,53].

The vanishing/exploding gradient problem makes it hard to learn and tune the parameters of the earlier layers in the network, which impedes convergence from the beginning. This results in the inability of models with deep layers to learn on a given dataset. The network performance with deep layers becomes saturated or even starts degrading rapidly.

The ResNet model was introduced to overcome these issues. The main idea behind ResNet models is to use residual blocks to improve the accuracy of the models. Residual blocks are based on the concept of “identity shortcut connection” that skips/bypasses one or more layers. The input of the residual block is denoted by  $x$ , and the output is  $H(x)$ , which is the desired underlying mapping. The difference or residual between them is shown in Equation (1):

$$F(x) = \text{Output} - \text{Input} = H(x) - x \quad (1)$$

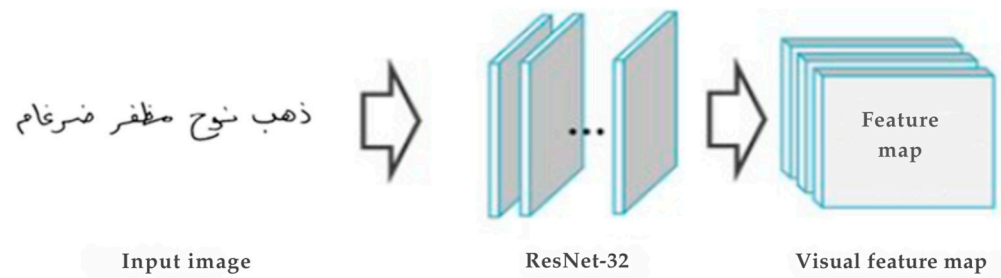
where  $F(x)$  is the mapping of the stacked nonlinear layers. The original mapping is rearranged into  $H(x) = F(x) + x$ . The additional  $x$  operates as a residual, thus “residual block”. Therefore, ResNet solves degradation by adding the input of a layer to its output. As a result, ResNet improves the efficiency of deep neural networks with more layers and avoids poor accuracy as the model becomes deeper.

In our model, we build a 32-layer ResNet-based model to extract character features [54]. The details of the network are illustrated in Table 3. The convolutional layers in the table are shown in the following format: (kernel size, stride (width)  $\times$  stride (height), pad (width)  $\times$  pad (height), channels). The max-pooling layers are shown in the following format: (kernel size, stride (width)  $\times$  stride (height), pad (width)  $\times$  pad (height)). The residual blocks in the ResNet model are shown in Table 3 with a gray background having the following format: [kernel size, channels]. Each convolution layer in the residual blocks has stride 0 and zero padding.

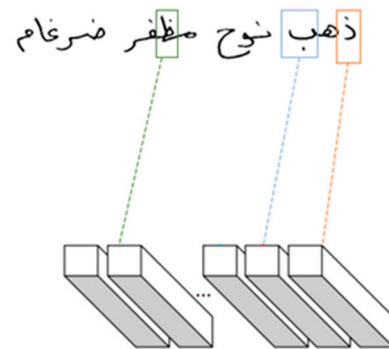
**Table 3.** ResNet model architecture.

Layers	Configurations	Output
Conv1	$3 \times 3, 1 \times 1, 1 \times 1, 16$ $3 \times 3, 1 \times 1, 1 \times 1, 32$	$64 \times 1048$
Conv2	Pool 1: $2 \times 2, 2 \times 2, 0 \times 0$ $\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 1$ $3 \times 3, 1 \times 1, 1 \times 1, 64$	$32 \times 524$
Conv3	Pool 2: $2 \times 2, 2 \times 2, 0 \times 0$ $\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$ $3 \times 3, 1 \times 1, 1 \times 1, 128$	$16 \times 262$
Conv4	Pool 3: $2 \times 2, 1 \times 2, 1 \times 0$ $\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 5$ $3 \times 3, 1 \times 1, 1 \times 1, 256$	$8 \times 263$
Conv5	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 3$ $2 \times 2, 1 \times 2, 1 \times 0, 256$ $2 \times 2, 1 \times 1, 0 \times 0, 256$	$3 \times 263$

The ResNet model is trained from scratch. The input of this stage is the normalized image. The output is a visual feature map containing each character’s characteristic features in the image, as shown in Figures 9 and 10.



**Figure 9.** ResNet model for text feature extraction.



**Figure 10.** Feature map for characters.

### 3.4. Classification

In this stage, after the features are extracted and the feature map, which contains the qualities and characteristics of a sequence of characters, is produced, a classifier is used to generate characters with the help of predefined classes. Since we are dealing with long sentences that contain a sequence of characters and words, we used Bi-directional Long-Short-Term Memory (BiLSTM) to capture the context information of the sentence.

The BiLSTM model was proposed by Graves and Schmidhuber [55] and is robust as a classifier and in sequence recognition models in different natural-language processing (NLP) tasks such as speech recognition [56], natural language understanding [57], machine translation [58], and sentiment analysis [59].

The BiLSTM model consists of two LSTMs to process sequence information in two directions. The first is taking the sequence of inputs in the forward direction (past to future). The other is taking the sequence of inputs in the backward direction (future to past). Therefore, BiLSTM will efficiently extract the full-text context information since it has access to the previous and following context. Thus, we have used BiLSTM in our Arabic handwritten text recognition system.

Moreover, multiple BiLSTMs can be stacked together to have a deep BiLSTM model. The deep BiLSTM model allows a higher level of abstraction in data than a shallow model. Having a deep BiLSTM model improved the performance of the speech recognition system [60].

After the last BiLSTM layer, each column of contextual features is mapped to an output label. The Connectionist Temporal Classification (CTC) output layer, which was proposed by Graves et al. [61], is adopted to predict the probability of an output label sequence. The CTC layer has several character outputs and one additional output known as 'blank'. The additional 'blank' output is helpful to avoid making decisions in uncertain zones, that is, in a low context area, instead of being trained to constantly predict a character.

The Arabic language has 28 letters, and each letter has one to four forms. We added the 28 Arabic letters with their different forms for each letter, number, and punctuation mark and ended up with a class size of 135.

The probability given by CTC is defined as for input sequence  $Y$ , where  $Y = y_1, \dots, y_T$ , and  $T$  is the length of the sequence, the output is the probability of  $\pi$ , which is defined in Equation (2) as:

$$p(\pi|Y) = \prod_{t=1}^T y_{\pi_t}^t \quad (2)$$

where  $y_{\pi_t}^t$  is the probability of having a character  $\pi_t$  at time step  $t$  [62]. A sequence-to-sequence mapping function  $M$  is defined on the sequence  $\pi$ . The mapping function  $M$  maps  $\pi$  onto  $l$ , where  $l$  is the final prediction output, by removing the repeated characters first, and then removing the blanks. For example,  $M$  maps “--مم-ث-ا-ل-ل-” onto “مثال”, where “-” represents blank. The conditional probability is defined as the total sum of probabilities of all  $\pi$  that are mapped by  $M$  onto  $Y$ , as shown in Equation (3).

$$p(l|Y) = \sum_{\pi: M(\pi)=l} p(\pi|Y) \quad (3)$$

### 3.5. Language Model

To improve recognition accuracy, LMs are used in many NLP models, such as handwritten text recognition and speech recognition. In our system, we used an  $n$ -gram language model, which is a statistical language modeling technique. The  $n$ -gram language model is a probabilistic model that predicts the probability of a sequence of words in a text.

The  $n$ -gram language models are simple in their structure, easy to calculate the word occurrence probability, and work best with high performance when trained on large amounts of data. In this work, a 3-gram language model was trained on the training corpus of the KHATT and AHTID/MW datasets. KenLM language model toolkit [63] was used to build the 3-gram language model. The KenLM toolkit is faster and uses less memory than other existing toolkits such as SRI Language Modeling [64] and IRST Language Modeling [65], improving system runtime performance.

## 4. Results and Discussions

### 4.1. System Settings and Parameters

We used Python 3 and PyTorch tools and libraries to implement our model. The code was implemented using Amazon Web Services. We used Amazon Elastic Compute Cloud with a 16GB NVIDIA V100 GPU.

The network configuration of our model is shown in Table 3. We used the architecture of the ResNet model to construct 32 trainable layers, which are a combination of convolutional layers with a ReLU (Rectified Linear Unit) activation function, max pooling layers with  $2 \times 2$  filters, and batch normalization layers. It is beneficial to add the batch normalization technique for training our intense neural network. The batch normalization layers have the effect of stabilizing the learning process and accelerating the training process of the neural network. Our system applied a dropout layer after the ResNet model with a dropout ratio of 0.2. The second dropout layer is applied after the BiLSTM layers with a dropout ratio of 0.2.

Dropout, a stochastic regularization technique, is applied in our neural network. The dropout technique helps prevent overfitting and reduce interdependent learning amongst the neurons in neural networks by dropping out units (i.e., neurons) from the neural network during the training process.

The output of the ResNet model, which contains the extracted sequence of visual features from normalized text line images, is fed into the BiLSTM model with 512 hidden units to generate the contextual sequence. Different depths of BiLSTM layers are used to compare the performance of our model when adding more bidirectional LSTM layers. The first experiment was performed using 2 layers of BiLSTM, and the second experiment was performed on 3 layers of BiLSTM. The BiLSTM network is followed by the CTC decoder to translate the contextual feature sequence to the character sequence. The CTC decoder has 135 output units to generate characters and predict words.

Finally, we added a 3-gram language model to our system to improve the recognition accuracy. The KenLM toolkit, a fast and memory-efficient language model, is used to build a 3-gram language model. The language model compares the weights assigned by CTC and LM. The predicted word with the highest weight will be replaced.

For optimization, we adopted the Adadelata optimizer, which is a robust learning rate method that does not require the manual setting of a learning rate. We set the training batch size to 24, and all images were scaled to 1048x64 in both training and testing. The data were split to 80% for training, 10% for validation, and the remaining 10% for testing. The description of parameters used for training is mentioned in Table 4. The overfitting is evaluated using the EarlyStopping method based on validation loss value.

**Table 4.** Parameters for the training process.

Parameter	Value
Batch size	24
Epoch	300
Learning rate	1.0
Optimizer	Adadelata
EarlyStopping	Validation loss
Dropout rate	0.2

#### 4.2. Performance Evaluation

The performance of handwriting recognition systems was evaluated in terms of WER and CER. We used these two metrics to assess the performance of our system. The WER and CER are based on the concept of Levenshtein edit distance, which is the minimum number of edit operations required to transform the output text into the ground truth text. The editing operations are substitutions, insertions, and deletions necessary to convert the source string into the reference string. The WER is calculated as follows:

$$WER = \frac{S_w + I_w + D_w}{N_w} \times 100 \quad (4)$$

where  $S$  is the total number of substituted words,  $I$  is the total number of inserted words,  $D$  is the total number of deleted words, and  $N$  is the total number of words in the evaluation set.

The CER is calculated as follows:

$$CER = \frac{S_c + I_c + D_c}{N_c} \times 100 \quad (5)$$

where  $S$  is the total number of substituted characters,  $I$  is the total number of inserted characters,  $D$  is the total number of deleted characters, and  $N$  is the total number of characters in the evaluation set.

#### 4.3. Experimental Results

The last stage of developing a handwriting recognition system is testing the system. This process used scaled images as input to the Arabic handwritten text recognition system. Two different datasets of Arabic handwritten text, that is, the KHATT and AHTID/MW, were used to cover all forms of Arabic text. Therefore, characters and words with different forms and widths were used in our experiments.

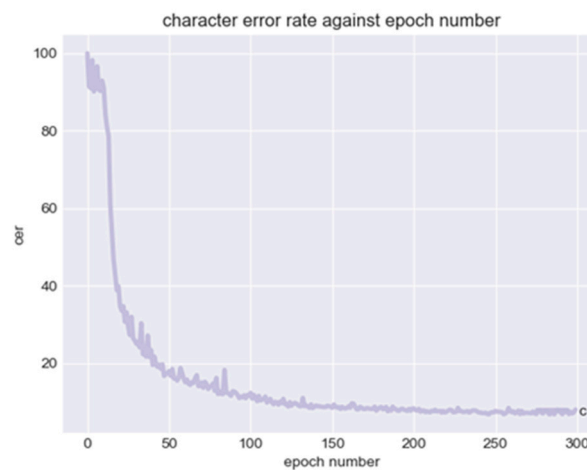
The scaled images were passed through the ResNet model, followed by the BiLSTM-CTC layers and the language model post-processing stage. Table 5 shows the results of our system using the KHATT and AHTID/MW datasets with different BiLSTM layers.

The recognition rates are improved in both datasets when using three BiLSTM layers. The WER is reduced by 4.29% for the KHATT dataset and by 5.37% for the AHTID/MW dataset. Therefore, the proposed model performs better, and the results are improved when using 3 layers of the BiLSTM network.

**Table 5.** Results of our system for the two datasets.

Dataset	Model	CER%	WER%
KHATT	2-BiLSTM Layers	15.8	31.6
	3-BiLSTM Layers	13.2	27.31
AHTID/MW	2-BiLSTM Layers	7.4	22.79
	3-BiLSTM Layers	6.6	17.42

Additional experiments were performed on the AHTID/MW dataset to test our proposed system performance. As seen below, the best performance is obtained by using 3-BiLSTM layers, which results in a 17.42% WER and 6.6% CER. Figures 11 and 12 show the relation between the CER and WER with the epoch number, respectively. As shown in Figures 11 and 12, the CER and WER decrease as the epoch number increases during the training process until it reaches epoch number 300. The results of our proposed system on the KHATT and AHTID/MW datasets confirm the robustness of our system.

**Figure 11.** CER versus epoch number.**Figure 12.** WER versus epoch number.

To validate our system's performance, we compared our results with the most recent works on Arabic handwriting recognition systems. Table 6 shows the results of recent works obtained from the test set of the KHATT and AHTID/MW datasets. The experimental results showed that our system had an impressive recognition accuracy with a WER of 27.31% and a CER of 13.2% on the test set of the KHATT corpus. ResNet is a resilient CNN architecture designed specifically for extracting information from textual images. The

fundamental concept underlying ResNet models is employing residual blocks to enhance the precision of the models. Residual blocks rely on the idea of an “identity shortcut connection” that allows for the skipping or bypassing of one or more levels.

**Table 6.** Comparison between our proposed system and the existing systems.

Reference	Year	Database	Model	CER	WER
BenZeghiba et al. [25]	2015	KHATT Dataset	MDLSTM	-	31.3%
Stahlberg and Vogel [27]	2015	KHATT Dataset	Kaldi’s network model	-	30.5%
Zeghiba [66]	2017	KHATT Dataset	MDLSTM	-	34.3%
Jemni et al. [24]	2018	KHATT Dataset	BiLSTM	16.27%	29.13%
Jemni et al. [67]	2019	AHTID/MW Dataset	CNN-MDLSTM	-	18.13%
Momeni [34]	2024	KHATT Dataset	Transformer	18.45%	-
Proposed Model	2024	KHATT Dataset	ResNet-BiLSTM-CTC	13.2%	27.31%
		AHTID/MW Dataset		6.6%	17.42%

## 5. Conclusions

We proposed a model for recognizing Arabic handwritten text. The system aims to identify Arabic handwritten text accurately by imitating the human brain to recognize text using machine learning approaches. The ResNet model was used for feature extraction, and BiLSTM-CTC sequence modeling was used for classification. Machine learning techniques are used to overcome traditional methods based on shallow learning and hand-engineered features. Moreover, machine learning approaches help overcome the challenges of recognizing Arabic handwritten text. A 3-gram language model was used in our system using the KenLM toolkit to improve the recognition accuracy of handwritten text.

Our proposed model was evaluated on the KHATT and AHTID/MW datasets. The experimental results showed that our system had an impressive recognition accuracy, with a 27.31% WER and a 13.2% CER for the KHATT dataset and a 17.42% WER and a 6.6% CER for the AHTID/MW dataset.

Although our proposed methodology for Arabic OCR has been successful, there are still certain limitations. The proposed study only employs a pretrained CNN model. Evaluating other transfer learning or transformer models is a future enhancement. Also, in future work, different datasets can be combined to reduce the generalizability problem.

**Author Contributions:** Conceptualization, methodology, supervision, and writing—review and editing, A.M.M.; data acquisition, methodology, experiment, and writing—original draft preparation, M.Y.A. and M.A.-H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are openly available [19,20].

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Nomenclature

### Symbols

$D_c$	Total number of deleted characters
$D_w$	Total number of deleted words
$F$	The mapping of the stacked nonlinear layers
$H$	The output of the residual block
$I_c$	Total number of inserted characters
$I_w$	Total number of inserted words
$l$	Final prediction output
$M$	Mapping function



$N_c$	Total number of characters in the evaluation set
$N_w$	Total number of words in the evaluation set
$p$	Probability
$S_c$	Total number of substituted characters
$S_w$	Total number of substituted words
$T$	Length of the sequence
$t$	Time step
$x$	The input of the residual block
$Y$	Input sequence
$\pi$	Sequence
Acronyms	
AHCD	Arabic Handwritten Character Dataset
AHTID/MW	The Arabic Handwritten Text Images Database written by Multiple Writers
AIA9K	AlexU Isolated Alphabet
ANN	Artificial Neural Network
AWS	Amazon Web Services
BiLSTM	Bidirectional Long Short-Term Memory
CER	Character Error Rate
CNN	Convolutional Neural Network
CTC	Connectionist Temporal Classification
DC	Distribution-Concavity
DeepAHR	Deep Neural Network for Arabic Handwritten Recognition
DPI	Dots Per Inch
ECR	Error Classification Rate
HACDB	Handwritten Arabic characters database
HMM	Hidden Markov Model
IESK-ArDB	Institute for Electronics, Signal Processing and Communications Arabic Database
KFUPM	The King Fahd University of Petroleum and Minerals
KHATT	KFUPM Handwritten Arabic Text Database
LM	Language Model
MDLSTMs	Multi-Dimensional Long Short-Term Memory Networks
NLP	Natural Language Processing
OCR	Optical Character Recognition
PAW	Part-of-Arabic-Words
PR	Pattern Recognition
ReLU	Rectified Linear Unit
ResNet	Residual Network
SHG	Segmentation Hypothesis Graph
SVM	Support Vector Machine
VGG	Visual Geometry Group
WER	Word Error Rate

## References

1. Chaudhuri, A.; Mandaviya, K.; Badelia, P.; Ghosh, S.K. Optical Character Recognition Systems. In *Optical Character Recognition Systems for Different Languages with Soft Computing*; Springer International Publishing: Cham, Switzerland, 2017; pp. 9–41.
2. Eberhard, D.M.; Simons, G.F.; Fennig, C.D. *Gujarati. Ethnologue: Languages of the World*, 22nd ed.; SIL International: Dallas, TX, USA, 2019.
3. Nashif, M.H.H.; Miah, M.B.A.; Habib, A.; Moulik, A.C.; Islam, M.S.; Zakareya, M.; Ullah, A.; Rahman, M.A.; Al Hasan, M. Handwritten numeric and alphabetic character recognition and signature verification using neural network. *J. Inf. Secur.* **2018**, *9*, 209. [\[CrossRef\]](#)
4. El-Dabi, S.S.; Ramsis, R.; Kamel, A. Arabic character recognition system: A statistical approach for recognizing cursive typewritten text. *Pattern Recognit.* **1990**, *23*, 485–495. [\[CrossRef\]](#)
5. Faizullah, S.; Ayub, M.S.; Hussain, S.; Khan, M.A. A Survey of OCR in Arabic Language: Applications, Techniques, and Challenges. *Appl. Sci.* **2023**, *13*, 4584. [\[CrossRef\]](#)
6. Anis, M.; Maalej, R.; Elleuch, M. Recent advances of ML and DL approaches for Arabic handwriting recognition: A review. *Int. J. Hybrid Intell. Syst.* **2023**, *19*, 61–78. [\[CrossRef\]](#)
7. AlKhateeb, J.H.; Jiang, J.; Ren, J.; Ipson, S. Component-based segmentation of words from handwritten Arabic text. *Int. J. Comput. Syst. Sci. Eng.* **2009**, *5*.

8. Nashwan, F.; Rashwan, M.A.; Al-Barhamtoshy, H.M.; Abdou, S.M.; Moussa, A.M. A holistic technique for an Arabic OCR system. *J. Imaging* **2018**, *4*, 6. [\[CrossRef\]](#)
9. Boufenar, C.; Kerboua, A.; Batouche, M. Investigation on deep learning for off-line handwritten Arabic character recognition. *Cogn. Syst. Res.* **2018**, *50*, 180–195. [\[CrossRef\]](#)
10. Alrobah, N.; Albahli, S. Arabic Handwritten Recognition Using Deep Learning: A Survey. *Arab. J. Sci. Eng.* **2022**, *47*, 9943–9963. [\[CrossRef\]](#)
11. Berriche, L.; Alqahtani, A.; RekikR, S. Hybrid Arabic handwritten character segmentation using CNN and graph theory algorithm. *J. King Saud Univ.—Comput. Inf. Sci.* **2024**, *36*, 101872. [\[CrossRef\]](#)
12. Mosbah, L.; Moalla, I.; Hamdani, T.M.; Neji, B.; Beyrouthy, T.; Alimi, A.M. ADOCRNet: A Deep Learning OCR for Arabic Documents Recognition. *IEEE Access* **2024**, *12*, 55620–55631. [\[CrossRef\]](#)
13. Mahdi, M.G.; Sleem, A.; Elhenawy, I. Deep Learning Algorithms for Arabic Optical Character Recognition: A Survey. *Multicriteria Algorithms Appl.* **2024**, *2*, 65–79. [\[CrossRef\]](#)
14. Ralaibozaka, T.C.N.; Rafidison, M.A.; Ramafiarisona, H.M. Contribution to the Authenticity of Digitized Handwritten Signatures Through Deep Learning with Resnet-50 and Ocr. *Int. J. Innov. Eng. Res. Technol.* **2024**, *11*, 20–25. [\[CrossRef\]](#)
15. Kayabas, A.; Topcu, A.E.; Kiliç, Ö. OCR Error Correction Using BiLSTM. In Proceedings of the 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), Cape Town, South Africa, 9–10 December 2021; pp. 1–5. [\[CrossRef\]](#)
16. Li, W.; Zhang, L.-C.; Wu, C.-H.; Wang, Y.; Cui, Z.-X.; Niu, C. A data-driven approach to RUL prediction of tools. *Adv. Manuf.* **2024**, *12*, 6–18. [\[CrossRef\]](#)
17. Graves, A. Connectionist Temporal Classification. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 61–93.
18. He, Y. Research on Text Detection and Recognition Based on OCR Recognition Technology. In Proceedings of the 2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE), Dalian, China, 27–29 September 2020; pp. 132–140. [\[CrossRef\]](#)
19. Mahmoud, S.A.; Ahmad, I.; Alshayeb, M.; Al-Khatib, W.G.; Parvez, M.T.; Fink, G.A.; Märgner, V.; El Abed, H. Khatt: Arabic offline handwritten text database. In Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition, Bari, Italy, 18–20 September 2012; pp. 449–454.
20. Mezghani, A.; Kanoun, S.; Khemakhem, M.; El Abed, H. A database for arabic handwritten text image recognition and writer identification. In Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition, Bari, Italy, 18–20 September 2012; pp. 399–402.
21. Mamouni El, M. An Effective Combination of Convolutional Neural Network and Support Vector Machine Classifier for Arabic Handwritten Recognition. *Autom. Control Comput. Sci.* **2023**, *57*, 267–275. [\[CrossRef\]](#)
22. Alheraki, M.; Al-Matham, R.; Al-Khalifa, H. Handwritten Arabic Character Recognition for Children Writing Using Convolutional Neural Network and Stroke Identification. *Hum.-Centric Intell. Syst.* **2023**, *3*, 147–159. [\[CrossRef\]](#)
23. Elleuch, M.; Maalej, R.; Kherallah, M. A new design based-SVM of the CNN classifier architecture with dropout for offline Arabic handwritten recognition. *Procedia Comput. Sci.* **2016**, *80*, 1712–1723. [\[CrossRef\]](#)
24. Jemni, S.K.; Kessentini, Y.; Kanoun, S.; Ogier, J.-M. Offline Arabic handwriting recognition using BLSTMs combination. In Proceedings of the 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), Vienna, Austria, 24–27 April 2018; pp. 31–36.
25. BenZeghiba, M.F.; Louradour, J.; Kermorvant, C. Hybrid word/Part-of-Arabic-Word Language Models for arabic text document recognition. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 671–675.
26. Forney, G.D. The viterbi algorithm. *Proc. IEEE* **1973**, *61*, 268–278. [\[CrossRef\]](#)
27. Stahlberg, F.; Vogel, S. The qcri recognition system for handwritten arabic. In Proceedings of the International Conference on Image Analysis and Processing, Genoa, Italy, 7–11 September 2015; pp. 276–286.
28. Povey, D.; Zhang, X.; Khudanpur, S. Parallel training of deep neural networks with natural gradient and parameter averaging. *arXiv* **2014**, arXiv:1410.7455.
29. Wigington, C.; Stewart, S.; Davis, B.L.; Barrett, W.A.; Price, B.L.; Cohen, S.D. Data Augmentation for Recognition of Handwritten Words and Lines Using a CNN-LSTM Network. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; pp. 639–645.
30. Altwaijry, N.; Al-Turaiqi, I. Arabic handwriting recognition system using convolutional neural network. *Neural Comput. Appl.* **2021**, *33*, 2249–2261. [\[CrossRef\]](#)
31. El Khayati, M.; Kich, I.; Taouil, Y. CNN-based Methods for Offline Arabic Handwriting Recognition: A Review. *Neural Process. Lett.* **2024**, *56*, 115. [\[CrossRef\]](#)
32. AlShehri, H. DeepAHR: A deep neural network approach for recognizing Arabic handwritten recognition. *Neural Comput. Appl.* **2024**, *36*, 12103–12115. [\[CrossRef\]](#)
33. Alghyaline, S. Optimised CNN Architectures for Handwritten Arabic Character Recognition. *Comput. Mater. Contin.* **2024**, *79*, 4905–4924. [\[CrossRef\]](#)
34. Momeni, S.; BabaAli, B. A transformer-based approach for Arabic offline handwritten text recognition. *Signal Image Video Process.* **2024**, *18*, 3053–3062. [\[CrossRef\]](#)

35. Mahmoud, S.A.; Ahmad, I.; Al-Khatib, W.G.; Alshayeb, M.; Parvez, M.T.; Märgner, V.; Fink, G.A. KHATT: An open Arabic offline handwritten text database. *Pattern Recognit.* **2014**, *47*, 1096–1112. [[CrossRef](#)]
36. Ahmad, R.; Naz, S.; Afzal, M.Z.; Rashid, S.F.; Liwicki, M.; Dengel, A. Khatt: A deep learning benchmark on arabic script. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; pp. 10–14.
37. Kaur, S. Noise types and various removal techniques. *Int. J. Adv. Res. Electron. Commun. Eng. (IJARECE)* **2015**, *4*, 226–230.
38. Soille, P. *Morphological Image Analysis: Principles and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
39. Stahlberg, F.; Vogel, S. Detecting dense foreground stripes in Arabic handwriting for accurate baseline positioning. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 361–365.
40. Tavoli, R.; Keyvanpour, M.; Mozaffari, S. Statistical geometric components of straight lines (SGCSL) feature extraction method for offline Arabic/Persian handwritten words recognition. *IET Image Process.* **2018**, *12*, 1606–1616. [[CrossRef](#)]
41. Mohamad, R.A.-H.; Likforman-Sulem, L.; Mokbel, C. Combining slanted-frame classifiers for improved HMM-based Arabic handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 1165–1177. [[CrossRef](#)]
42. Akram, H.; Khalid, S. Using features of local densities, statistics and HMM toolkit (HTK) for offline Arabic handwriting text recognition. *J. Electr. Syst. Inf. Technol.* **2017**, *4*, 387–396.
43. Jayech, K.; Mahjoub, M.A.; Amara, N.E.B. Arabic handwriting recognition based on synchronous multi-stream HMM without explicit segmentation. In Proceedings of the International Conference on Hybrid Artificial Intelligence Systems, Bilbao, Spain, 22–24 June 2015; pp. 136–145.
44. Benouareth, A.; Ennaji, A.; Sellami, M. Semi-continuous HMMs with explicit state duration for unconstrained Arabic word modeling and recognition. *Pattern Recognit. Lett.* **2008**, *29*, 1742–1752. [[CrossRef](#)]
45. Almodfer, R.; Xiong, S.; Mudhsh, M.; Duan, P. Multi-column deep neural network for offline Arabic handwriting recognition. In Proceedings of the International Conference on Artificial Neural Networks, Alghero, Italy, 11–14 September 2017; pp. 260–267.
46. Zhao, Z.-Q.; Zheng, P.; Xu, S.-t.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [[CrossRef](#)]
47. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 818–833.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
49. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [[CrossRef](#)]
50. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. [[CrossRef](#)]
51. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
52. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [[CrossRef](#)]
53. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS), Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010; pp. 249–256.
54. Cheng, Z.; Bai, F.; Xu, Y.; Zheng, G.; Pu, S.; Zhou, S. Focusing attention: Towards accurate text recognition in natural images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5076–5084.
55. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [[CrossRef](#)] [[PubMed](#)]
56. Graves, A.; Jaitly, N.; Mohamed, A.-r. Hybrid speech recognition with deep bidirectional LSTM. In Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 8–12 December 2013; pp. 273–278.
57. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S.R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv* **2018**, arXiv:1804.07461.
58. McCann, B.; Bradbury, J.; Xiong, C.; Socher, R. Learned in translation: Contextualized word vectors. *arXiv* **2017**, arXiv:1708.00107.
59. Chen, T.; Xu, R.; He, Y.; Wang, X. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Syst. Appl.* **2017**, *72*, 221–230. [[CrossRef](#)]
60. Graves, A.; Mohamed, A.-r.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6645–6649.
61. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, New York, NY, USA, 29 June 2006; pp. 369–376.

62. Shi, B.; Bai, X.; Yao, C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 2298–2304. [[CrossRef](#)] [[PubMed](#)]
63. Heafield, K. KenLM: Faster and smaller language model queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation, Edinburgh, UK, 30–31 July 2011; pp. 187–197.
64. Stolcke, A.; Zheng, J.; Wang, W.; Abrash, V. SRILM at sixteen: Update and outlook. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, Waikoloa, HI, USA, 11–15 December 2011; p. 5.
65. Federico, M.; Bertoldi, N.; Cettolo, M. IRSTLM: An open source toolkit for handling large scale language models. *Interspeech* **2008**, 1618–1621.
66. Zeghiba, M.F.B. Arabic word decomposition techniques for offline Arabic text transcription. In Proceedings of the 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), Nancy, France, 3–5 April 2017; pp. 31–35.
67. Jemni, S.K.; Kessentini, Y.; Kanoun, S. Out of vocabulary word detection and recovery in Arabic handwritten text recognition. *Pattern Recognit.* **2019**, *93*, 507–520. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.