Revolutionizing Historical Document Digitization: LSTM-Enhanced OCR for Arabic Handwritten Manuscripts

Safiullah Faizullah[1], Muhammad Sohaib Ayub[2], Turki Alghamdi[3], Toqeer Syed Ali[4],
Muhammad Asad Khan[5], Emad Nabil[6]
Faculty of Computer and Information Systems, Islamic University of Madinah, KSA[1,3,4,6]
Department of Computer Science, Lahore University of Management Sciences, Pakistan[2]
Department of Telecommunication, Hazara University, Pakistan[5]

***Abstract*—Optical Character Recognition (OCR) holds immense practical value in the realm of handwritten document analysis, given its widespread use in various human transactions. This scientific process enables the conversion of diverse documents or images into analyzable, editable, and searchable data. In this paper, we present a novel approach that combines transfer learning and Arabic OCR technology to digitize ancient handwritten scripts. Our method aims to preserve and enhance accessibility to extensive collections of historically significant materials, including fragile manuscripts and rare books. Through a comprehensive examination of the challenges encountered in digitizing Arabic handwritten texts, we propose a transfer learning-based framework that leverages pre-trained models to overcome the scarcity of labeled data for training OCR systems. The experimental results demonstrate a remarkable improvement in the recognition accuracy of Arabic handwritten texts, thereby offering a highly promising solution for the digitization of historical documents. Our work enables the digitization of large collections of ancient historical materials, including manuscripts and rare books characterized by delicate physical conditions. The proposed approach signifies a significant step towards preserving our cultural heritage and facilitating advanced research in historical document analysis.**

Keywords*—Optical character recognition; transfer learning; Arabic OCR; image processing; classification; convolutional neural network

I. Introduction

Around 1.8 billion people in the world speak the Arabic language. Arabic writing is unique and semi-cursive in both printed and handwritten forms. Arabic OCR systems are of two types, online and offline, aiming to convert Arabic text images into machine-readable words. Online systems use special equipment like a pen and tablet, while offline systems use scanners. There are open issues in Arabic OCR, such as generalization ability, the use of deep learning, lack of standard taxonomy, large-scale evaluation, and reproducible research [1], [2], [3].

The method used to process the documents or images to extract text is represented as OCR. These images or documents be in different forms, like scanned or digitized. This process helps extract text from these documents in an editable form for a machine to edit it. This process exists in two types, i.e. online and offline. In an online OCR system, real-time text recognition will be performed like writing on digital notebooks [4]. Whereas, written documents or images taken from some written sources are used in the offline OCR system [5]. As we focus on Arabic and dealing with handwritten documents or images, it comes in both offline and online types. Therefore, recognizing the text more accurately is the issue because the Arabic language has unique and challenging characteristics compared to other languages [6]. Its characters come in different shapes concerning their position in words, increasing the difficulty level for recognition, and when the document is of low quality or in different writing styles, fonts, cursive nature, and quality documents [7].

Some available tools help extract and recognize the Arabic text from the documents, like Tesseract [8], OCR Space [9], OmniPage [10], easy-OCR [11] and others. From all these, Tesseract gives a better result, but it needs more training and data to recognize handwritten Arabic text with high character, word, and overall text accuracy. By keeping this point in mind, we decided to use the transfer learning method on Tesseract to achieve a high recognition rate for offline handwritten Arabic text. It needs more training datasets, including ground truth, images, and box files. For this purpose, we write a script that makes images of the ground truth to fast forward the process and then makes the box files, and then these files are used together for transfer learning.

OCR can be used in many fields of life where it makes work easy, efficient, and digital. Hospitals process patient and insurance company files on the computer to make a digitized record as these kinds of records are handwritten, which is why OCR needs handwritten documents as well and uses them in many more fields [12], [13], [14].

A. Problem Formulation

Many OCR systems are available for different languages with different features, like unilingual or multilingual OCR systems. Much research is done in English and other Germanic languages, especially in noncursive scripts, because these are easier to be processed by the OCR systems. However, scripts like Arabic are challenging to recognize due to their cursive nature, the appearance of Arabic letters in different shapes in different words as shown in Table I, and diacritics that come above or below the Arabic word or letter. It changes the meaning of words as they are just minor signs, as shown in Table II. Arabic script is written from right to left, and some writing

styles change the appearance of the words, like in Fig. 1. Due to these complexities, Arabic OCR is a challenging research area that needs more consideration. Limited publicly available datasets exist for research purposes, and different techniques are available. However, each technique has pros and cons that limit character and word accuracy in the preprocessing and segmentation phases [15], [16].

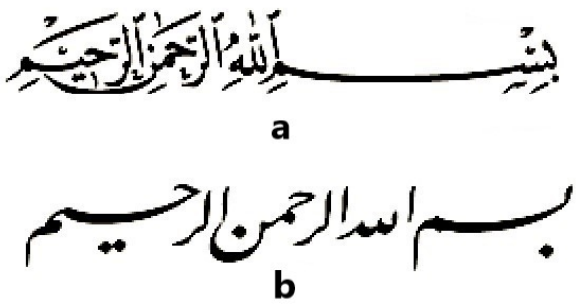

Fig. 1. Arabic sentence in two different writing styles, i.e. (a) Nask and (b) Nastaliq.

Table I. Arabic Letters and their Different Shapes based on the Position in a Word

Isolated Form	Initial Form	Middle Form	End Form
ا	ا	ـا	ـا
ب	بـ	ـبـ	ـب
ت	تـ	ـتـ	ـت
ث	ثـ	ـثـ	ـث
ج	جـ	ـجـ	ـج
ح	حـ	ـحـ	ـح
خ	خـ	ـخـ	ـخ
د	د	ـد	ـد
ذ	ذ	ـذ	ـذ
ر	ر	ـر	ـر
ز	ز	ـز	ـز
س	سـ	ـسـ	ـس
ش	شـ	ـشـ	ـش
ص	صـ	ـصـ	ـص
ض	ضـ	ـضـ	ـض
ط	طـ	ـطـ	ـط
ظ	ظـ	ـظـ	ـظ
ع	عـ	ـعـ	ـع
غ	غـ	ـغـ	ـغ
ف	فـ	ـفـ	ـف
ق	قـ	ـقـ	ـق
ک	کـ	ـکـ	ـک
ل	لـ	ـلـ	ـل
م	مـ	ـمـ	ـم
ن	نـ	ـنـ	ـن
ه	هـ	ـهـ	ـه
و	و	ـو	ـو
ي	يـ	ـيـ	ـي

B. Problem Motivation

The digitization of historical documents stands as a critical endeavor in the preservation and dissemination of our global cultural heritage. These documents, ranging from ancient manuscripts to letters and administrative records, encapsulate not only the factual history but also the intellectual, social, and cultural dynamics of past societies. Digitization offers a bulwark against the relentless march of time, safeguarding these irreplaceable insights from the ravages of physical degradation, environmental hazards, and the obscurity of inaccessibility. Moreover, it democratizes access to knowledge, transcending geographical and temporal barriers to make these treasures universally accessible, fostering a broader understanding and appreciation of human history. In the realm of Arabic historical documents, which are rich in linguistic and cultural nuances, digitization is not merely a technical challenge but a crucial step towards preserving a significant part of the world's intangible cultural heritage. Thus, our research is motivated by the imperative to advance Optical Character Recognition (OCR) technologies, specifically tailored to Arabic script, to facilitate the efficient and accurate digitization of these documents, ensuring their preservation and accessibility for future generations.

Table II. Position of Arabic Diacritics in the Script Changes the Meaning and Sound of the Words

Diacritics Type	Diacritics Shape
Fathah	◌َ
Kasrah	◌ِ
Dammah	◌ُ
Alif Khanjariyah	◌ٰ
Sukūn	◌ْ
Tanwin	◌ً ◌ٌ ◌ٍ
Shaddah	◌ّ

The motivation for this research paper on Arabic handwritten OCR stems from the challenges associated with recognizing and digitizing handwritten Arabic documents. While there are existing OCR systems for recognizing printed Arabic text, there has been less research on developing accurate models for recognizing handwritten Arabic characters and numbers. This is problematic because handwritten documents are important cultural artifacts that need to be preserved, but they are at risk of being lost if not digitized. Furthermore, the lack of publicly available datasets for Arabic handwritten OCR makes it difficult for researchers to develop and evaluate new models. Therefore, this paper aims to contribute to the field of Arabic handwritten OCR by presenting a new dataset and an accurate OCR model that can be used to preserve and make accessible handwritten Arabic documents.

We make the following contributions to significantly advance the state-of-the-art in Optical Character Recognition (OCR) for Arabic handwritten texts:

1) We present an enhanced OCR accuracy for Arabic handwritten texts through the novel application of transfer learning techniques to the Tesseract OCR engine, substantially reducing common recognition errors.
2) Our work includes the compilation and preparation of a comprehensive dataset, which comprises high-resolution `.tiff` images, `.box` files, `ground truth` files, and `dictionary files`. This dataset not only supports our model's training but also serves as a resource for the broader research community.
3) We introduce a robust framework for evaluating OCR accuracy, utilizing Character Error Rate (CER) and Word Error Rate (WER) as the principal metrics. This framework facilitates a thorough quantitative analysis and validation of our approach.
4) Through detailed **visual demonstrations of our**

model's efficacy, we provide clear evidence of our methodology's effectiveness and its practical implications for Arabic text digitization.

These contributions underscore the novelty and significance of our work, showcasing its potential to significantly impact the OCR field by enhancing the recognition accuracy of Arabic handwritten texts and promoting the digitization and preservation of historical documents.

II. Related Work

The exploration of Optical Character Recognition (OCR) technologies, especially in the realm of Arabic handwritten texts, reveals a landscape marked by both advancements and persistent challenges. While the strides in OCR methodologies have paved the way for notable achievements, the unique intricacies of Arabic script—ranging from its cursive nature to the prevalence of diacritics—pose specific hurdles that remain inadequately addressed. This gap not only underscores the necessity for innovative approaches but also serves as the cornerstone of our motivation. Our investigation into the related work illuminates the breadth of strategies previously employed, yet simultaneously highlights a critical void in the application of advanced machine learning techniques, such as transfer learning, to the nuances of Arabic handwriting recognition. It is this intersection of opportunity and challenge that our research aims to navigate, propelled by the conviction that enhancing OCR accuracy for Arabic texts not only contributes to the technological domain but also fosters the preservation and accessibility of cultural heritage. As such, our work seeks not only to bridge the existing gaps identified through our review but to set a new benchmark for accuracy and efficiency in the recognition of Arabic handwritten documents.

The Arabic language has complexities for developing OCR systems due to its cursive nature and morphological structure. Applied approaches and methods in this area are compared and show the best-performing approach. This study shows that the hidden Markov model (HMM) gives an accuracy of 95.6%. As the survey is working on handwritten characters recognition of Arabic, it shows the recognition by an automatic capitalization of OCR using the hidden Markov model [17].

Arabic text recognition system faces many challenges due to its cursive nature, different shapes of characters, diacritics, and writing styles [7], [18]. Handwritten text recognition comes in two formats, i.e. document (manuscripts) and online (tablets) [19]. A proper pipeline is required to overcome the challenges and get a high recognition rate [20]. Urdu and Arabic are the same, i.e. Urdu uses all of the Arabic characters. Recent advances discuss state-of-the-art research about Urdu and Arabic Naskh and Nastaliq scripts. It has also discussed the dataset and its different forms, i.e. printed, scanned, and handwritten, and the pipeline used to process the data, i.e. preprocessing, segmentation, classification, recognition, and post-processing to get better text recognition [21], [22], [23].

As mentioned above, most researchers also work on Arabic and use transfer learning. However, they only concentrated on characters and numbers, which are easier to recognize through transfer learning on models like Alex-Net and Google-Net. Moreover, they are already trained on the image of different objects and give high recognition on a single letter or number images [24].

The authors of [25] discuss the challenges of handwritten Arabic text classification and recognition and evaluate the different deep learning models, i.e. ResNet50, ResNet101, VGG16, VGG19, AlexNet, GoogleNet, and ResNet18 using transfer learning techniques. To overcome the challenges, use handwritten images of text written by a native or non-native person. The dataset consists of 22 subjects equally written by native and non-native writers. After using different models, results show that GoogleNet is the model that achieves 93.2% accuracy on the native dataset and 95.2% on the non-native dataset.

Sahlol et al. [26] describe a hybrid machine-learning approach for handwritten Arabic character recognition using optical character recognition (OCR) systems. This approach combines neighborhood rough sets with a binary whale optimization algorithm for feature selection. The proposed method outperformed state-of-the-art and deep neural networks regarding recognition rate and computational time. However, some misclassified failure cases occurred due to the context of appearance.

Authors of [27] present a new model for Arabic document information retrieval (ADIR) using OCR services. They used datasets written by 60 writers containing 16,800 Arabic letters and applied them to preprocess techniques such as binarization, noise removal, skew correction, and page and zone segmentation. They discussed the challenges of segmentation and recognition for Arabic script, particularly with cursive writing and compound graphemes. The Arabic document information retrieval (ADIR) system achieved a classification success rate of 100% for test images. The paper also describes the service description for ADIR, which includes the user interface (UI) and a server address for communication between clients and services.

Rahal et al. [28] discussed the difficulties of automatic Arabic text recognition due to the language's cursive nature, character similarities, large vocabulary, and the use of multiple font sizes. This paper proposed a novel hybrid network that combines Hidden Markov Models (HMMs) and a Bag-of-Feature (BoF) framework, which is based on a deep Sparse Auto-Encoder (SAE) for feature extraction. The system eliminates the need for preprocessing stages like baseline estimation and slant normalization. Instead, the SAE automatically selects the best weights for visual words for each local descriptor while concurrently learning the best dictionary, making it suitable for irregular, variable-size, mixed-font, high- and low-resolution documents. The system was tested on four different datasets and performed well on each one.

Zanona et al. [29] present a model for recognizing handwritten Arabic characters that use preprocessing functions and contour analysis to produce a vector for recognition by a neural network. The system was tested on private data and achieved 98% accuracy on the complete dataset and 99.4% precision. The classification system uses a segmentation operation [30] and a multilayer feed-forward neural network (FFNN). In another research, an Arabic handwritten text recognition system was designed that extracts and recognizes single-line text and converts that extracted or predicted text into individual words

and their characters, achieving an 83% recognition rate.

Many advanced systems, like Mathpix, Digital Ink API, ML Kit Text, Read-Ink, MyScript, GoodNotes, and Mazec, were developed in the last decade to resolve challenges in handwritten text recognition. Multiple methods and techniques are used for this purpose, like dynamic time warping, hidden Markov models, and artificial neural networks, which use a pipeline to process the data correctly. These methods use the data and enhance the recognition rate, which shows impressive results [31].

Some systems are used for the Arabic text OCR, i.e. Tesseract, Textract, and Document AI. These OCR systems work for different languages, like English and Arabic. In this research, historical documents are used, and these documents are multilingual and contain English and Arabic text. The English dataset contains historical text from books that scan with various fonts. In contrast, the Arabic dataset contains articles from online resources in a single widely used font. The author found that Textract and Tesseract performed slightly better on gray-scale test images than on color versions, but Tesseract was more sensitive to noise than the other two engines. It shows that Document AI and Textract give better results than Tesseract. However, Document AI and Textract have higher noise in their images even after applying noise removal technique in it [32], which recommends self-training or transfer learning to the Tesseract, it will be able to give better results than others.

Most of the researchers work in Arabic handwritten OCR. However, they mostly used characters, digits, or other individuals in their experiments and showed better results as shown in Table III.

Authors in [35] proposed a technique based on text area detection and text recognition using pre-trained OCR systems, i.e. Tesseract, KerasOCR, and EasyOCR. This system deals with engineering documents, and a high recognition rate is significant. For this purpose, transfer learning is used, which helps to increase the word recognition rate and increase the overall text recognition rate as well.

III. Methodology

This section describes the methodology of our proposed work. We provide a descriptive analysis of our datasets. We have explained our transfer learning approach and each step of the experimentation.

A. Dataset

As datasets play an important role in training models, a more extensive dataset helps train the model more accurately, and the model can learn quickly. For deep and machine learning models, larger datasets are required, but due to handwritten text, data preparation is a tough job as first need to generate round truth. Handwritten text against ground truth and its image and box files generation to pass the model for text extraction, recognition, and further processing. Due to these challenges, the availability of a public dataset is very low. Moreover, every model requires a large amount of data to do its excellent training; by considering this problem, we moved toward transfer learning, and we know that for transfer learning models, a larger dataset is not an issue, and this technique helps the model to get better training even on smaller datasets.

Urdu and Arabic have many similarities, i.e. both are written from right to left, all of the Arabic characters are used in Arabic, but Urdu has just extra characters, their digits have the same shapes, their writing styles are also common, and Urdu borrows a large amount of vocabulary from Arabic that is about 30%.

Due to these similarities, the Urdu dataset can also be used for Arabic text. Bhatti et al. [36] presents an Urdu dataset based on the Urdu handwritten digits dataset and experiments using deep learning techniques and achieving a high accuracy rate.

An Urdu handwritten dataset is proposed by [37], which is based on Nastaliq handwritten text, i.e. UNHD, and used a bidirectional LSTM classifier and this dataset is written by 500 writers on A4 size paper as shown in Table IV that will be available on request. The address of the dataset link is attached [1].

Mostafa et al. [38] proposed a dataset that is based on text images shown in Fig. 2 with their ground truth as well. The dataset contains 270 million words and 1.6 billion characters. 12 fonts are used in the dataset, and the used text font size is 13.

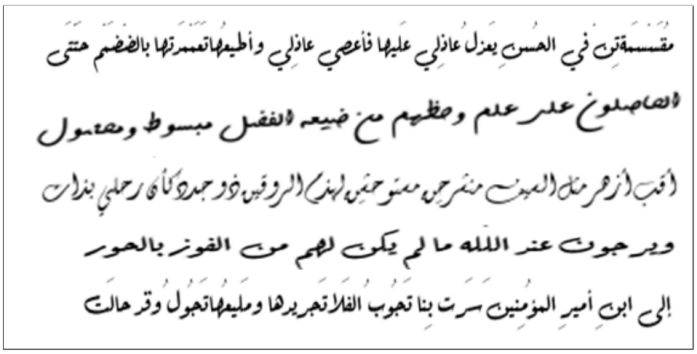

Fig. 2. Dataset sample of arabic text image that contains text with diacritics.

The authors of [39] proposed a handwritten dataset that is taken from different handwritten documents. A description of the dataset is shown in Table V and a sample image of the dataset is shown in Fig. 3.

Dataset selection:

For dataset selection, we find the Arabic text and then write a Python script that tests a script that gives the relative frequency of characters present in the text, and then we compare the relative frequency to that relative frequency [40], [41], as shown in the Table VI to avoid biases, this frequency distribution helps to use the characters in same quantity and position as they are used in normal text writing. Also, we test the script on each file separately to check their frequency. We then gave the text files to different people to write the handwritten text, and then we scanned that text and evaluated the handwritten dataset.

[1] https://sites.google.com/site/researchonurdulanguage1/databases

Table III. Previous Arabic Handwritten Text Research Focused on Character Recognition, Digit Recognition, Hijja Recognition, and Word Recognition using Image-based Data for Model Training

Paper	Model	Dataset	Accuracy
Albhattah [33]	Hybrid CNN with finetune	AHCD (characters and Hijja)	98%
	CNN with finetune	AHCD (characters)	92.4%
Alghyaline [34]	Deep CNN	APTI (words)	76.30%
	CNN-RNN	Alif dataset (words)	85.98%
	Deep CNN	HMBD (characters)	92.88%

Table IV. Description of the Urdu Nastaliq Handwritten Dataset, adaptable for Arabic Due to its Derivation from the Arabic Script

Urdu Nastaliq Handwritten Dataset	Description
Writers	500
Text lines	10,000
Words	312,000
Characters	1,872,000
Words written by a writer	624

Table V. Description of the Handwritten Dataset Extracted from Handwritten Documents

No. of Pages	No. of Lines	No. of Words	No. of Chars
1,000	18,000	35,000	252,000

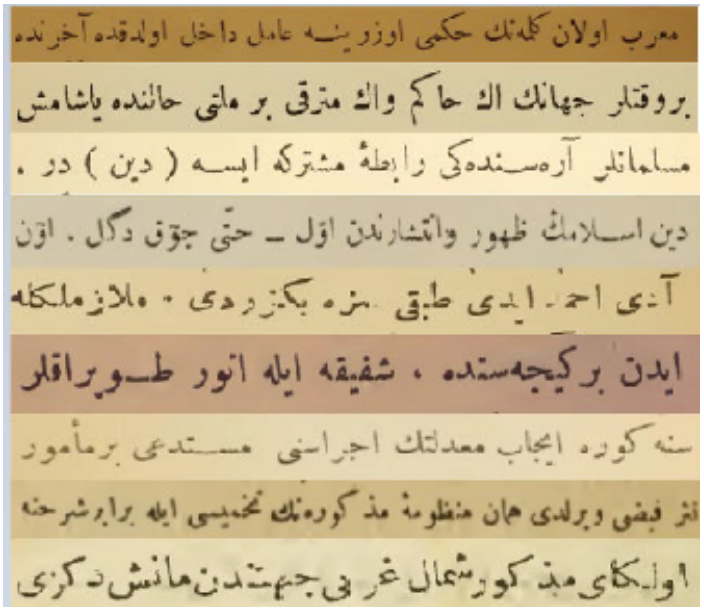

Fig. 3. Sample image of the handwritten dataset that is taken from the various Arabic handwritten documents.

1) Publicly Available Datasets: Some handwritten datasets are reported [2], but those datasets contain images of words because most researchers work on individual character and number recognition of Arabic. They used these datasets for pre-trained object image recognition models like ImageNet and AlexNet [19]. A description of these kinds of datasets is shown in Table VII.

Arabic MNIST dataset is also available, but it is also for individual characters and numbers. This database is split into training and testing datasets for Arabic characters and numbers, consisting of 60,000 images for training and 10,000 images for testing purposes of characters and the same for numbers; this dataset is available [3]. Furthermore, another dataset is available that is written by the contribution of 60 writers and contains 168,00 character images, and these images were scanned on 300dpi resolution [42].

Table VI. Comparison of Arabic Characters Frequency in the Selected Text and Overall Frequency of Arabic Characters

Characters	Frequency in selected text	Frequency in Arabic Language
ا	15.00%	14.61%
ل	14.13%	11.64%
م	6.56%	6.49%
ي	6.38%	7.25%
و	5.64%	5.40%
ن	5.41%	4.76%
ت	4.39%	4.58%
ع	4.15%	3.27%
ر	3.54%	4.53%
ب	2.70%	3.38%

Table VII. Description of Handwritten Images of Arabic Words Dataset used in Pre-trained Models

Statistics	AlexU-W	IFN/ENIT
Images	25,114	32,492
Training images	20,114	26,459
Testing images	5,029	6,033
Unique words	109	937
Maximum PAWS	3	10

A gold-standard dataset is available on GitHub [4], which contains Arabic books, their ground truth, images, and box files. That is much enough to train any model, text images based on one-line text, and a description of the gold standard set is shown in Table VIII. Some publicly available datasets of Arabic handwritten text, words, characters, and digits are also available, as shown in Table IX.

Table VIII. EDA of Arabic Gold-Standard Dataset Derived from Literature Books

Book	Pages	Lines	Words	Chars
IbnFaqihHamadhani.al-Buldan	79	1466	16909	92730
IbnAthir.al-kamil	40	794	12818	58481
Ibn Qutayba.Adab al-katib	55	794	7848	42230
al-Jahiz.al-Hayawan	65	992	11870	59191
al-Yacqubi.al-Tarikh	68	1050	13487	66341
al-Dhahabi.Tarikh al-islam	50	1110	11045	55047
Ibn al-Jawzi.al-Muntazam	50	938	13156	62574

We used a publicly available dataset for our experiments, which consists of images of handwritten text, ancient Arabic dataset, and printed Arabic text data for comparison and its ground truth as well [52], [53]. The dataset consists of about

[2]http://www.eng.alexu.edu.eg/~mehussein/alexu-word/

[3]http://yann.lecun.com/exdb/mnist/

[4]https://github.com/OpenArabic/OCR_GS_Data/tree/master/ara

Table IX. Publicly Available Datasets Focus Primarily on Arabic Ligatures and Digits

Dataset	Type of content
IFN/ENIT [43]	Handwritten Words
HACDB [44]	Handwritten Characters
KHATT [45]	Handwritten Text lines
SmartATID [46]	Printed & Handwritten Pages
Degraded historical [47]	Handwritten documents
Numeral [48]	Handwritten Digits
AHDBIFTR [49]	Handwritten images
ARABASE [50]	Handwritten Text
CENPARMI [51]	Handwritten subwords & digits

25,000 entries, which contain ground truth, images, box, and "lstmf" files, which is a complete set of the required dataset for training the Tesseract transfer learning, and these images consist of one-line text, and in a raw format that will be preprocessed in the preprocessing stage of Tesseract pipeline, as shown in Table X.

Table X. EDA of Datasets used for Evaluating Handwritten Arabic Text in Our Experiments

Dataset	Words	Char	Lines	Digits	Punc.
Printed lines	3,592	19,802	307	29	370
Words and lines	71,365	374,516	14,606	5,592	6,451
IbnFaqihHamadhani	104,845	577,392	15,296	1,713	9,948
Ancient Arabic	700	3,160	100	-	-

B. Transfer Learning using Tesseract

Tesseract is an open-source engine for Optical Character Recognition used to recognize text from images. It consists of the following steps to process the image from its raw shape to make it able to be used for OCR, i.e. preprocessing, Converting Image to Box file, Converting Box files to "lstmf" file.

While using the Tesseract base model, firstly, we passed the handwritten image to the model to check its accuracy and try to find the problem in the base model. After passing images in PNG or JPG, we write a script that takes all the images from the folder individually, passes them from the model, extracts text from the images, and saves them in a text file in another folder. After this, we pass both ground truth and extracted or predicted text files and find accuracy. For evaluation, we write a script using Levenshtein distance to find the character error rate (CER) and word error rate (WER) and then find the average CER and WER and their accuracy as the base model is trained on editable or computer-typed text. Therefore, we prepare the data for model [5] some dictionary files, the dataset contains ".tiff" files and ground truth and passed this data to the Tesseract for transfer learning purpose and consider starting point to the base model. After transfer learning, repeat the test for evaluation and get the results with more accuracy, CER and WER, as shown in Fig. 4.

The methodology depicted in Fig. 4 commences with the meticulous preparation of training data, a pivotal phase for the effective employment of transfer learning on the Tesseract OCR engine. This stage encompasses the assembly of high-resolution `.tiff` images of Arabic handwritten texts, alongside their corresponding `.box` files which delineate bounding boxes around each character, and `ground truth` files containing the verbatim text. Furthermore, `dictionary files` are curated to guide the model towards recognizing the expected lexical items. Subsequent to the data preparation, the Tesseract base model undergoes augmentation through transfer learning, enabling it to adapt to the nuances of Arabic handwritten text. This adaptation is facilitated by training the model with the prepared dataset, thereby enhancing its capability to accurately recognize Arabic characters and words. The retrained model is then evaluated against a set of unseen data to ascertain its accuracy, employing metrics such as the Character Error Rate (CER) and Word Error Rate (WER) derived through Levenshtein distance calculations. These metrics serve as indicators of the model's proficiency in character and word recognition, with lower CER and WER values signifying superior performance. This comprehensive approach underscores the efficacy of transfer learning in refining the Tesseract model's recognition accuracy for Arabic handwritten texts.

Preprocessing: In this step, Tesseract takes an image for preprocessing and improves the image quality by resizing the image into a standard size, converting the image into gray-scale, and applying filters to remove noise from the image like thresholding, erosion, blur.

Converting image into Box file: In this step, it converts the preprocessed image into a box file with plain text containing coordinates of the bounding boxes around each character or word of text present in an image. The following command is used to make a box file against the image. "Tesseract image.png output -l ara makebox". In this command, "image.png" is an image that needs to be recognized, and 'output' is the file that gets the recognized text and saves it, while "ara" is for language; here, we are using Arabic, and we used the keyword for it, same "eng" is for the English language.

Converting box file to "lstmf": This step takes the box file created in the previous step and makes a binary file containing the training data for Tesseract. This process contains the following steps like "tesseract image.png output -l ara box.train", Which will create "lstmf" files with the name of "ara.traineddata" in the same directory.

Evaluation: For evaluating, we need to compare the recognized text with the ground truth to get accuracy. For this purpose, we can use the following steps to get a better evaluation, i.e. CER, WER, and Overall accuracy. A higher accuracy, lower CER, and BER represent good results. CER gives the percentage of incorrectly recognized characters. It can be calculated simply by dividing the number of incorrect characters by the number of characters in the ground truth, as shown in Eq. (1) below.

$$\text{CER} = \frac{\text{Number of incorrect characters}}{\text{Total number of characters}} \tag{1}$$

WER gives the percentage of incorrectly recognized words. It can be calculated simply by dividing the number of incorrect words by the total number of words in the ground truth, as shown in Eq. (2) below.

$$\text{WER} = \frac{\text{Number of incorrect words}}{\text{Total number of words}} \tag{2}$$

[5] https://github.com/Shreeshrii/tesstrain-JSTORArabic/tree/master/data/

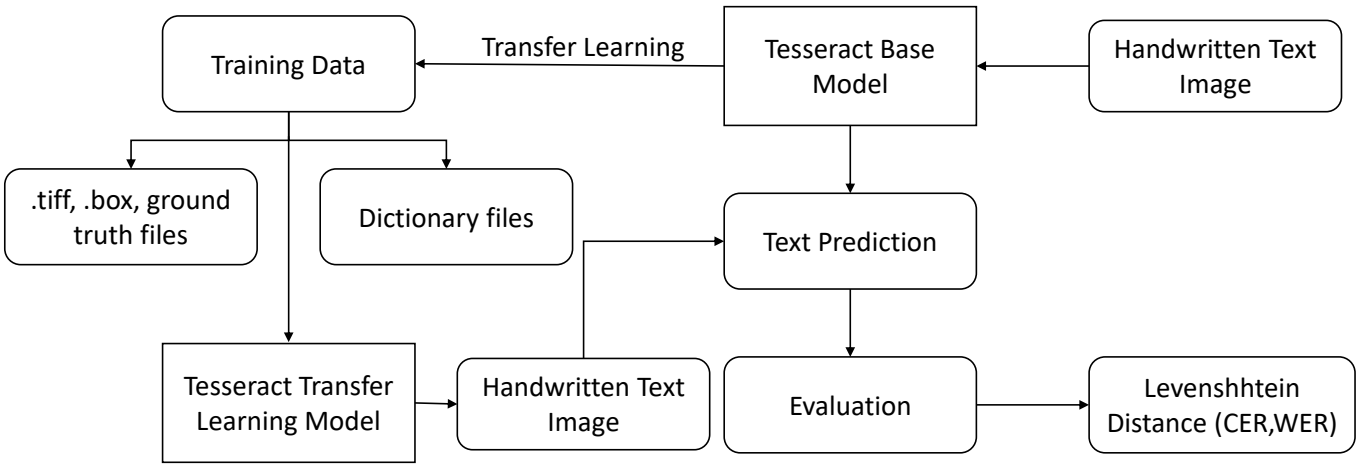

Fig. 4. Tesseract base-model, transfer learning process and evaluation methods.

Overall accuracy gives the percentage of correctly recognized words and characters. It can be calculated by dividing the number of correct words and characters by the total number of words and characters in the ground truth, as shown in Eq. (3) below.

$$\text{Accuracy} = \frac{\text{No. of correct } \{\text{words + characters}\}}{\text{Total no. of } \{\text{words + characters}\}} \quad (3)$$

IV. Results

The experiments were conducted on a Windows Dell laptop with an Intel(R) Core(TM) i7-6600U CPU @ 2.60GHz, 16GB DDR4 RAM, 512GB SSD storage, and an Intel® HD Graphics 520. The laptop ran Windows 10 and was connected to a stable power source throughout the experiments. The experiments were conducted in a controlled environment to minimize external factors that could affect the results. Additionally, Ubuntu 22.04.1 LTS was installed using WSL to run the Tesseract experiment.

In the next section, evaluation metrics, i.e. CER and WER, are discussed, and the overall accuracy of the predicted text against ground truth is described with the results. Then, a discussion of the results and experiments is presented.

A. Evaluation Metrics

After transfer learning of Tesseract, we evaluate the model. For this purpose, firstly, give handwritten text image files to the model that generates box and "lstmf" file and then gives the predicted text of all images. Then, this predicted text compares with ground truth, which is also available in a text file. As mentioned above, the dataset is split into 80% for the training and 20% for testing. Furthermore, we used two types of evaluation, i.e. our evaluation that gives us the overall accuracy of the experiment, CER and WER. For more detail, we use an open-source evaluation tool named "OCREVALUATION", which compares ground truth and the predicted text that elaborates more openly, as shown in Fig. 5. This sample image has two sections: one represents the ground truth on the left, and the other is about the predicted text. The ground truth section represents the characters or words by different colors that are predicted wrong in the predicted text section.

The base model gives an overall accuracy of 23.30%, an average CER of 31.57%, and an average WER of 65.95% on handwritten text images of the Arabic language. However, after transfer learning, it gives an overall accuracy of 87.89% and gives the average CER of 14.02% and average WER of 41.39%, which is relatively better, also shown in Table XI. It also depends on the size of the dataset, and the total time taken by the training is 21 hours. Then, evaluate the second dataset, which contains 5526 images of text that are also based on 1 line text. For these images, we also have ground truth for evaluation purposes and then generate "lstmf" and box files for each image and the ground truth file. After evaluation, we get a character error rate (CER) of 14.85%, a word error rate (WER) of 40.30%, and achieve an overall accuracy of 85.53% also shown in Table XI.

Table XI. Comparison of Tesseract base Model with our Transfer Learning-based Model for Arabic Handwritten Text Recognition

Model	CER	WER	Accuracy%
Tesseract Base Model	31.57%	65.95%	23.30%
Transfer Learning (Dataset 1)	14.02%	41.39%	87.89%
Transfer Learning (Dataset 2)	14.85%	40.30%	85.53%

To check the accuracy of the printed text of this retrained model, we test a dataset that contains a total of 6118 files, which are divided into image and ground truth files, and that took about 5 hours to evaluate. It achieves an accuracy of 94.94%, character error rate, and word error rate based on order dependent and independent are shown in Table XII.

Table XII. Arabic Printed Images Text Recognition with Order Dependent and Order Independent

Features	Stats
Number of files	3,059
CER (order independent)	5.62%
WER(order independent)	19.08%
CER (order dependent)	5.55%
WER(order dependent)	17.85%

Fig. 5. Spotting differences between ground truth (gt.txt) and predicted (ara.OCR.txt) arabic handwritten text.

B. Results and Discussion

For this problem, we have chosen the open-source OCR engine Tesseract as the base model and applied the transfer learning approach to get better Arabic handwritten text recognition results. Firstly, we tested the base model to know its results on Arabic text; we decided to evaluate the computer-generated Arabic text first. After evaluation, we found an average CER of 14.02% and an average WER of 41.39% and got an overall accuracy of about 87.89%. Then, we tested the handwritten images on the base model and got bad results, giving an average CER of 31.57%, WER of 65.95%, and accuracy of 23.30%.

Some challenges while recognizing text are that Arabic handwriting can vary significantly between individuals, making it difficult for Tesseract to recognize characters accurately. This variation can be due to factors such as writing style, speed of writing, and individual handwriting quirks. Arabic script includes diacritical marks, which are symbols that indicate vowel sounds. These marks can be challenging for Tesseract to recognize accurately, especially when small or poorly written.

Handwritten text may contain noise or distortion due to uneven ink distribution, smudging, or poor image quality. This can make it difficult for Tesseract to recognize characters accurately. Handwritten text may be oriented or aligned in various ways, making it challenging for Tesseract to recognize the correct characters or words. Additionally, training Tesseract with a more extensive and diverse dataset of Arabic handwriting may improve its accuracy.

After this, we apply the transfer learning approach and train the model by giving handwritten images and their ground truth. After 21 hours of training, the transfer learning gives outstanding results, with CER being 14.02%, WER being 41.39%, and overall accuracy of 87.89%, which is impressive. Then, evaluate the second dataset, which contains 5526 images of text that are also based on 1 line text. After evaluation, we get a character error rate (CER) of 14.85%, a word error rate (WER) of 40.30%, and achieve an overall accuracy of 85.53%.

V. Conclusion

In conclusion, our paper has presented an effective OCR method for handwritten Arabic text recognition using a transfer learning approach with Tesseract. Our evaluation results show a significant improvement in word and character recognition accuracy compared to previous models. The applied transfer learning technique achieved an average CER of 14.02% and an average WER of 41.39% and got an overall accuracy of about 87.89%. Then we decided to test the handwritten images on the base model to get the comparison result with the retrained model by using the transfer learning technique, giving an average CER of 31.57%, WER of 65.95%, and accuracy of 23.30%. These findings suggest that transfer learning can be a valuable technique for improving OCR accuracy in challenging languages such as Arabic and may provide a promising direction for future research in this field. Overall, our work demonstrates the potential of leveraging existing knowledge and models to improve the performance of OCR systems.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

Conceptualization: S.F., M.S.A., T.S.A., M.A.K. and A.A.; Data curation: M.S.A. and T.S.A.; Formal analysis: S.F.; Funding acquisition: S.F. and A.A.; Investigation: M.S.A., T.S.A. and M.A.K.; Methodology: M.S.A. and T.S.A.; Project administration: S.F. and M.A.K.; Resources: S.F. and A.A.; Software: M.S.A. and T.S.A.; Supervision: S.F. and M.A.K.; Validation: A.A.; Visualization: M.S.A. and T.S.A.; Writing – original draft: S.F., M.S.A., T.S.A., M.A.K. and A.A.; Writing - review & editing: S.F., M.S.A., T.S.A., M.A.K. and A.A.;

Acknowledgments

This work is funded by the Deputyship of Research & Innovation, Ministry of Education in Saudi Arabia, through project number 964. In addition, the authors would like to express their appreciation for the support provided by the Islamic University of Madinah.

References

[1] S. Djaghbellou, A. Bouziane, A. Attia, and Z. Akhtar, "A survey on arabic handwritten script recognition systems," *International Journal of Artificial Intelligence and Machine Learning (IJAIML)*, vol. 11, no. 2, pp. 1–17, 2021.

[2] L. S. Al-Homed, K. M. Jambi, and H. M. Al-Barhamtoshy, "A deep learning approach for arabic manuscripts classification," *Sensors*, vol. 23, no. 19, p. 8133, 2023.

[3] R. Najam and S. Faizullah, "Analysis of recent deep learning techniques for arabic handwritten-text ocr and post-ocr correction," *Applied Sciences*, vol. 13, no. 13, p. 7568, 2023.

[4] B.-G. Han, J. T. Lee, K.-T. Lim, and D.-H. Choi, "License plate image generation using generative adversarial networks for end-to-end license plate character recognition from a small set of real images," *Applied Sciences*, vol. 10, no. 8, p. 2780, 2020.

[5] A. F. d. S. Neto, B. L. D. Bezerra, and A. H. Toselli, "Towards the natural language processing as spelling correction for offline handwritten text recognition systems," *Applied Sciences*, vol. 10, no. 21, p. 7711, 2020.

[6] K. M. Nahar, I. Alsmadi, R. E. Al Mamlook, A. Nasayreh, H. Gharaibeh, A. S. Almuflih, and F. Alasim, "Recognition of arabic air-written letters: Machine learning, convolutional neural networks, and optical character recognition (ocr) techniques," *Sensors*, vol. 23, no. 23, p. 9475, 2023.

[7] S. Faizullah, M. S. Ayub, S. Hussain, and M. A. Khan, "A survey of ocr in arabic language: Applications, techniques, and challenges," *Applied Sciences*, vol. 13, no. 7, p. 4584, 2023.

[8] R. Smith, "An overview of the tesseract ocr engine," in *Ninth international conference on document analysis and recognition (ICDAR 2007)*, vol. 2. IEEE, 2007, pp. 629–633.

[9] T. C. Wei, U. Sheikh, and A. A.-H. Ab Rahman, "Improved optical character recognition with deep neural network," in *2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA)*. IEEE, 2018, pp. 245–249.

[10] F. Harbuzariu, C. Irimia, and A. Iftene, "Official document text extraction using templates and optical character recognition," in *2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA)*. IEEE, 2023, pp. 1–4.

[11] N. Awalgaonkar, P. Bartakke, and R. Chaugule, "Automatic license plate recognition system using ssd," in *2021 International Symposium of Asian Control Association on Intelligent Robotics and Industrial Automation (IRIA)*. IEEE, 2021, pp. 394–399.

[12] A. Kumar, P. Singh, and K. Lata, "Comparative study of different optical character recognition models on handwritten and printed medical reports," in *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*. IEEE, 2023, pp. 581–586.

[13] F. Azzam, M. Jaber, A. Saies, T. Kirresh, R. Awadallah, A. Karakra, H. Barghouthi, and S. Amarneh, "The use of blockchain technology and ocr in e-government for document management: Inbound invoice management as an example," *Applied Sciences*, vol. 13, no. 14, p. 8463, 2023.

[14] H. Butt, M. R. Raza, M. J. Ramzan, M. J. Ali, and M. Haris, "Attention-based cnn-rnn arabic text recognition from natural scene images," *Forecasting*, vol. 3, no. 3, pp. 520–540, 2021.

[15] S. Bergamaschi, S. De Nardis, R. Martoglia, F. Ruozzi, L. Sala, M. Vanzini, and R. A. Vigliermo, "Novel perspectives for the management of multilingual and multialphabetic heritages through automatic knowledge extraction: The digitalmaktaba approach," *Sensors*, vol. 22, no. 11, p. 3995, 2022.

[16] F. M. Nashwan, M. A. Rashwan, H. M. Al-Barhamtoshy, S. M. Abdou, and A. M. Moussa, "A holistic technique for an arabic ocr system," *Journal of Imaging*, vol. 4, no. 1, p. 6, 2017.

[17] A. S. Shaker, "A survey for an automatic transliteration of arabic handwritten script," *Journal of Physics: Conference Series*, vol. 1530, no. 1, p. 012094, 2020.

[18] P. Ahmed and Y. Al-Ohali, "Arabic character recognition: Progress and challenges," *Journal of King Saud University-Computer and Information Sciences*, vol. 12, pp. 85–116, 2000.

[19] M. Awni, M. I. Khalil, and H. M. Abbas, "Offline Arabic handwritten word recognition: A transfer learning approach," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 10, pp. 9654–9661, 2022.

[20] S. Naz, K. Hayat, M. I. Razzak, M. W. Anwar, S. A. Madani, and S. U. Khan, "The optical character recognition of urdu-like cursive scripts," *Pattern Recognition*, vol. 47, no. 3, pp. 1229–1248, 2014.

[21] A. M. Alshanqiti, S. Albouq, A. B. Alkhodre, A. Namoun, and E. Nabil, "Employing a multilingual transformer model for segmenting unpunctuated arabic text," *Applied Sciences*, vol. 12, no. 20, p. 10559, 2022.

[22] A. Qaroush, B. Jaber, K. Mohammad, M. Washaha, E. Maali, and N. Nayef, "An efficient, font independent word and character segmentation algorithm for printed arabic text," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 1, pp. 1330–1344, 2022.

[23] A. Alshanqiti, A. Namoun, A. Alsughayyir, A. M. Mashraqi, A. R. Gilal, and S. S. Albouq, "Leveraging distilbert for summarizing arabic text: an extractive dual-stage approach," *IEEE Access*, vol. 9, pp. 135 594–135 607, 2021.

[24] M. A. KO and S. Poruran, "OCR-nets: variants of pre-trained CNN for Urdu handwritten character recognition via transfer learning," *Procedia Computer Science*, vol. 171, pp. 2294–2301, 2020.

[25] A. A. Almisreb, S. Turaev, M. A. Saleh, S. A. M. Al Junid *et al.*, "Arabic Handwriting Classification using Deep Transfer Learning Techniques," *Pertanika Journal of Science & Technology*, vol. 30, no. 1, pp. 641–654, 2022.

[26] A. T. Sahlol, M. Abd Elaziz, M. A. Al-Qaness, and S. Kim, "Handwritten arabic optical character recognition approach based on hybrid whale optimization algorithm with neighborhood rough set," *IEEE Access*, vol. 8, pp. 23 011–23 021, 2020.

[27] H. M. Al-Barhamtoshy, K. M. Jambi, S. M. Abdou, and M. A. Rashwan, "Arabic documents information retrieval for printed, handwritten, and calligraphy image," *IEEE Access*, vol. 9, pp. 51 242–51 257, 2021.

[28] N. Rahal, M. Tounsi, A. Hussain, and A. M. Alimi, "Deep sparse auto-encoder features learning for arabic text recognition," *IEEE Access*, vol. 9, pp. 18 569–18 584, 2021.

[29] M. A. Zanona, A. Abuhamdah, and B. M. El-Zaghmouri, "Arabic hand written character recognition based on contour matching and neural network." *Comput. Inf. Sci.*, vol. 12, no. 2, pp. 126–137, 2019.

[30] A. Zoizou, A. Zarghili, and I. Chaker, "A new hybrid method for arabic multi-font text segmentation, and a reference corpus construction," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 5, pp. 576–582, 2020.

[31] T. Ghosh, S. Sen, S. M. Obaidullah, K. Santosh, K. Roy, and U. Pal, "Advances in online handwritten recognition in the last decades," *Computer Science Review*, vol. 46, p. 100515, 2022.

[32] T. Hegghammer, "OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment," *Journal of Computational Social Science*, vol. 5, no. 1, pp. 861–882, 2022.

[33] W. Albattah and S. Albahli, "Intelligent Arabic Handwriting Recognition Using Different Standalone and Hybrid CNN Architectures," *Applied Sciences*, vol. 12, no. 19, p. 10155, 2022.

[34] S. Alghyaline, "A Printed Arabic Optical Character Recognition System using Deep Learning," *Journal of Computer Science*, vol. 18, no. 11, pp. 1038–1050, 2022.

[35] W. Khallouli, R. Pamie-George, S. Kovacic, A. Sousa-Poza, M. Canan, and J. Li, "Leveraging Transfer Learning and GAN Models for OCR from Engineering Documents," in *World AI IoT Congress (AIIoT)*. IEEE, 2022, pp. 015–021.

[36] A. Bhatti, A. Arif, W. Khalid, B. Khan, A. Ali, S. Khalid, and A. u. Rehman, "Recognition and classification of handwritten urdu numerals using deep learning techniques," *Applied Sciences*, vol. 13, no. 3, p. 1624, 2023.

[37] S. B. Ahmed, S. Naz, S. Swati, and M. I. Razzak, "Handwritten urdu character recognition using one-dimensional blstm classifier," *Neural Computing and Applications*, vol. 31, no. 4, pp. 1143–1151, 2019.

[38] A. Mostafa, O. Mohamed, A. Ashraf, A. Elbehery, S. Jamal, A. Salah, and A. S. Ghoneim, "An end-to-end ocr framework for robust arabic-handwriting recognition using a novel transformers-based model and an innovative 270 million-words multi-font corpus of classical arabic with diacritics," *arXiv preprint arXiv:2208.11484*, 2022.

[39] İ. Dölek and A. Kurt, "A deep learning model for ottoman ocr," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 20, p. e6937, 2022.

[40] S. Boudelaa, M. Perea, and M. Carreiras, "Matrices of the frequency and similarity of arabic letters and allographs," *Behavior Research Methods*, vol. 52, pp. 1893–1905, 2020.

[41] Wikipedia, "Arabic letter frequency," https://www.intellaren.com/articles/en/a-study-of-arabic-letter-frequency-analysis, 2023, [Accessed 05-12-2023].

[42] A. El-Sawy, M. Loey, and H. El-Bakry, "Arabic handwritten characters recognition using convolutional neural network," *WSEAS Transactions on Computer Research*, vol. 5, no. 1, pp. 11–19, 2017.

[43] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze, H. Amiri *et al.*, "IFN/ENIT-database of handwritten Arabic words," in *Proc. of CIFED*, vol. 2. Citeseer, 2002, pp. 127–136.

[44] A. Lawgali, M. Angelova, and A. Bouridane, "HACDB: Handwritten Arabic characters database for automatic character recognition," in *European workshop on visual information processing (EUVIP)*. IEEE, 2013, pp. 255–259.

[45] S. A. Mahmoud, I. Ahmad, W. G. Al-Khatib, M. Alshayeb, M. T. Parvez, V. Märgner, and G. A. Fink, "KHATT: An open Arabic offline handwritten text database," *Pattern Recognition*, vol. 47, no. 3, pp. 1096–1112, 2014.

[46] F. Chabchoub, Y. Kessentini, S. Kanoun, V. Eglin, and F. Lebourgeois, "SmartATID: A mobile captured Arabic Text Images Dataset for multi-purpose recognition tasks," in *International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2016, pp. 120–125.

[47] A. Sulaiman, K. Omar, and M. F. Nasrudin, "A database for degraded Arabic historical manuscripts," in *International Conference on Electrical Engineering and Informatics (ICEEI)*. IEEE, 2017, pp. 1–6.

[48] S. M. Awaidah and S. A. Mahmoud, "A multiple feature/resolution scheme to Arabic (Indian) numerals recognition using hidden Markov models," *Signal Processing*, vol. 89, no. 6, pp. 1176–1184, 2009.

[49] J. Ramdan, K. Omar, M. Faidzul, and A. Mady, "Arabic handwriting data base for text recognition," *Procedia Technology*, vol. 11, pp. 580–584, 2013.

[50] N. E. B. Amara, O. Mazhoud, N. Bouzrara, and N. Ellouze, "ARABASE: A Relational Database for Arabic OCR Systems." *Int. Arab J. Inf. Technol.*, vol. 2, no. 4, pp. 259–266, 2005.

[51] Y. Al-Ohali, M. Cheriet, and C. Suen, "Databases for recognition of handwritten Arabic cheques," *Pattern Recognition*, vol. 36, no. 1, pp. 111–121, 2003.

[52] R. Najam and S. Faizullah, "A scarce dataset for ancient arabic handwritten text recognition," *Data in Brief*, vol. 56, p. 110813, 2024.

[53] R. Najam and Faizullah, "Historical arabic handwritten text recognition dataset, mendeley data," https://data.mendeley.com/datasets/xz6f8bw3w8/1, 2024, [Accessed 16-10-2024].