International Journal of Computers and Informatics

Journal Homepage: https://www.ijci.zu.edu.eg

Int. j. Comp. Info. Vol. 1 (2023) 9–20

Paper Type: Original Article

Towards Robust Arabic and Urdu OCR Systems: A Systematic Review of Deep Learning Techniques

Mohamed G. Mahdi [1,2,*], **Ahmed Sleem** [3], **Ibrahim M. Elhenawy** [4] **and Soha Safwat** [5]

[1] Department of Computer Science, Higher Institute for Computer Sciences and Information Systems, 5th Settlement, New Cairo, Egypt; mohamed.grisha@cis.edu.eg.

[2] Department of Computer Science, Faculty of Computers and Informatics, Zagazig University, Zagazig, Sharqiyah, Egypt; m.gresha@fci.zu.edu.eg.

[3] Department of Computer Science, Faculty of Computers and Informatics, Tanta University, Tanta, Egypt; Ahmed.selim@ics.tanta.edu.eg.

[4] Department of Computer Science, Faculty of Computers and Informatics, Zagazig University, Zagazig, Sharqiyah, Egypt; ielhenawy@zu.edu.eg.

[5] Department of Computer Science, Faculty of Computers and Information Systems, Egyptian Chinese University, Cairo, Egypt; ssafwat@ecu.edu.eg.

Received: 20 Jul 2023 **Revised:** 03 Nov 2023 **Accepted:** 28 Nov 2023 **Published:** 01 Dec 2023

Abstract

In recent years, deep learning has increasingly replaced traditional machine learning algorithms across various domains, including Machine Translation (MT), Pattern Recognition (PR), Natural Language Processing (NLP), Speech Recognition (SR), and Computer Vision (CV). Notably, deep learning-based systems for optical character recognition (OCR) have demonstrated substantial success. However, within the domains of pattern recognition and computer vision, handwritten character recognition remains one of the most complex challenges. This complexity arises from the variability in character height, orientation, and width, as individuals employ diverse writing instruments and exhibit distinct writing styles. As a result, handwritten recognition becomes a particularly difficult task. Additionally, research on regional languages such as Arabic and Urdu remains relatively underexplored. This article presents a review and comparative analysis of the most significant deep learning techniques employed in the recognition of Arabic-adapted scripts, specifically focusing on Arabic and Urdu languages.

Keywords: Arabic Natural Language Processing, Urdu Natural Language Processing, Optical Character Recognition, Handwritten Character Recognition, Deep Learning.

1 | Introduction

Optical Character Recognition (OCR) is a technology designed for the automated conversion of printed or handwritten text within scanned or photographed images into machine-readable text. Since its inception during the early stages of computing, OCR has experienced significant advancements, particularly driven by the progress in machine learning and computer vision methodologies. These continuous enhancements have greatly improved the accuracy and efficiency of OCR systems[1–3].

OCR technology operates by analyzing the visual features of an input image, focusing on characteristics such as the size, shape, and positioning of individual characters. Classification algorithms are then applied to accurately identify and convert each character into digital text. This technology is versatile and can be applied

Corresponding Author: m.gresha@fci.zu.edu.eg

to a wide range of document types, including receipts, books, newspapers, handwritten notes, and business cards [4].

The practical applications of OCR technology are broad and include automating data entry processes, digitizing historical archives, and improving accessibility for individuals with visual impairments. When integrated with other technologies like Natural Language Processing (NLP) and Machine Translation (MT), OCR enables more advanced functionalities, such as Sentiment Analysis (SA) and Automated Language Translation (ALT), significantly expanding its utility in various fields [5].

While OCR technology has revolutionized the processing and analysis of textual information, it is not without limitations, and several challenges persist in its application. A primary difficulty lies in handling the variations in the input images, such as differences in font style, size, or spacing, as well as inconsistencies in handwriting. For accurate results, OCR algorithms must be carefully designed and trained to account for these variations. Despite these challenges, the societal impact of OCR technology is expected to grow in the future, driven by ongoing advancements and improvements [6–8].

Deep learning has significantly advanced OCR technology, enabling unprecedented levels of accuracy in character recognition. Through the use of deep neural networks, OCR systems can effectively learn to identify characters and words, even in the presence of noise or image distortion. This has led to more reliable and precise recognition, particularly in challenging conditions where traditional OCR methods may struggle[9,10].

A key advantage of deep learning-based OCR systems is their ability to recognize characters and words by analyzing their visual features, rather than depending on predefined patterns or templates. This flexibility allows these systems to adapt more effectively to new fonts, writing styles, and languages, enhancing their applicability in diverse, real-world scenarios. Their adaptability makes them more robust in handling a variety of text formats and improving overall performance across different domains[11].

Deep learning-based OCR systems generally employ two main approaches: fully supervised learning and semi-supervised or unsupervised learning. In fully supervised learning, a large dataset of labeled images containing input characters or words is required to train the model, ensuring high accuracy and specificity. In contrast, semi-supervised or unsupervised learning can function with a smaller set of labeled images and leverage unlabeled data during training. This reduces the dependency on extensive labeled datasets while still enabling effective learning, making the latter approaches more efficient and scalable in scenarios where labeled data is limited [2,10].

Overall, deep learning has substantially enhanced the accuracy and efficiency of OCR systems, leading to significant advancements in character recognition technologies. These improvements have facilitated a wide range of new applications, including the digitization of documents and automatic text recognition within images or video surveillance [12]. By leveraging deep learning techniques, OCR systems are now more capable of handling complex recognition tasks in real-world scenarios, further expanding their practical utility across various industries[13,14].

The study of handwritten character recognition focuses on developing algorithms and methodologies for the automatic identification and conversion of handwritten characters into machine-readable or digital text. This field plays a crucial role in numerous sectors, including healthcare, education, and finance, which continue to depend significantly on handwritten documents. By improving the accuracy and efficiency of handwritten character recognition, this research area can enhance data accessibility and streamline workflows across these industries[15].

With advancements in Machine Learning and Computer Vision, the field of handwritten character recognition has seen remarkable progress in recent years. These systems typically begin by pre-processing the input image, followed by segmenting the characters into their components, and finally employing classification algorithms to recognize each character. Deep Learning techniques, particularly Convolutional Neural Networks (CNN), have proven highly effective in achieving cutting-edge performance in this area. However, challenges such as the variability of handwriting styles and the requirement for large annotated datasets remain significant hurdles. Despite these obstacles, continued progress in this field holds substantial potential for improving the efficiency and accuracy of processing handwritten documents, which is crucial in sectors that still rely heavily on manual documentation.

This work investigates various deep learning techniques applied to the recognition of Arabic-adapted scripts and Urdu. In this research, studies conducted over the past five years are highlighted, while also considering relevant survey papers in this domain. Additionally, the review encompasses studies from earlier years, providing a thorough overview of the field. The structure of this paper is as follows: Section 2 reviews prior studies that have employed deep learning for Arabic language recognition. In Section 3, we examine earlier research utilizing deep learning techniques for recognizing other languages, such as Urdu. Finally, Section 4 presents the conclusions drawn from this research.

2 |Arabic Natural Language Processing

Arabic is the most widely spoken language among the Semitic languages, serving as the official language in 26 countries and spoken by approximately 372 million individuals globally. Additionally, Arabic holds significant cultural and religious importance, as it is the language in which the Holy Quran is written, making it an essential medium for Muslims worldwide. As a result, Arabic ranks as the sixth most commonly spoken language today. Furthermore, it is one of the six official languages recognized by the United Nations (UN), alongside Chinese, English, French, Russian, and Spanish [16]. Arabic writing employs a cursive script and is read from right to left, utilizing 28 distinct characters. The size of each character is variable, influenced by factors such as the character's shape, the font used, and its position within a word (whether at the beginning, middle, end, or in isolation). The Arabic writing system also incorporates diacritical marks to denote short vowels and other phonetic elements. Examples of these diacritics include "fatha," "dhumma," "Tanween," and "kasra." Furthermore, Arabic features numerous ligatures formed by combining two or more characters, such as the "alif-laam" ligature [17,18].

Characters can also be distinguished by the presence of one to three dots, in addition to unique characters like "Hamza." Moreover, a single term in English may have multiple meanings when translated into Arabic, as exemplified by the words "good" and "love." Arabic has also influenced other languages, including Farsi, Kurdish, Urdu, and Pashto, which borrow words, characteristics, and structural elements from Arabic. These complexities present significant challenges for researchers in the field of Arabic language comprehension [19].

In [20], the authors present Hijja, a novel dataset comprising Arabic letters written by children aged 7 to 12, and introduce an automatic handwriting recognition model utilizing Convolutional Neural Networks (CNN). Their model is trained on both the Hijja dataset and the Arabic Handwritten Character Dataset (AHCD), achieving impressive accuracies of 97% and 88% for the AHCD and Hijja datasets, respectively. The authors further demonstrate that their model surpasses other existing models in the literature across all performance metrics. However, they acknowledge that the Hijja dataset poses greater challenges compared to the AHCD, as evidenced by the model's lower performance on the former. Overall, this study highlights the significant potential of convolutional neural networks in the recognition of Arabic handwriting.

In [21], the authors introduce a novel deep learning system named AHCR-DLS, designed for recognizing Arabic handwritten characters to tackle the inherent challenges associated with this task. The AHCR-DLS system achieves high accuracy rates on the training dataset, with average accuracies of 97.3% for HMB1 and 96.8% for HMB2. Similarly, on the test dataset, it demonstrates average accuracies of 95.5% for HMB1 and 94.9% for HMB2. These results underscore the effectiveness of the proposed approach in enhancing Arabic handwritten character recognition.

In [22], the authors propose a novel method for classifying handwritten Arabic characters utilizing a convolutional neural network (CNN) combined with an optimized leaky ReLU activation function. The results indicate that the proposed approach achieved high accuracy, surpassing both the standard ReLU (97.8%) and leaky ReLU (97.9%) activation functions across the four datasets employed in the study. Overall, this paper illustrates that the application of deep learning techniques can significantly enhance the performance of handwritten Arabic character recognition systems compared to traditional methods and other activation functions, such as ReLU and leaky ReLU.

Elkhayati and Elkettani [23] present a novel directed CNN model, referred to as UnCNN, designed for the recognition of isolated Arabic handwritten characters. The paper evaluates the effectiveness of this approach by comparing it with BsCNN and other contemporary models. The UnCNN model is assessed using four benchmark databases: IFHCDB, AHCD, AIA9K, and HACDB. The results indicate that UnCNN achieves competitive performance relative to some recent models in the literature, while outperforming others, thereby demonstrating its potential for effective Arabic handwritten character recognition.

In [24], Alghyaline introduces a novel approach for recognizing printed Arabic characters using deep learning, termed Printed Arabic Optical Character Recognition (PAOCR). This approach leverages the state-of-the-art You Only Look Once (YOLO) object detector and incorporates four key techniques. The first technique involves customizing and training YOLOv4 on deep Convolutional Neural Networks (CNNs) to enhance Arabic character recognition. The second technique processes overlapping bounding boxes to ensure the most accurate selection for each character. The third technique employs the Hunspell library to verify word spelling and correct any identified errors. Finally, the fourth technique utilizes edit distance to compare OCR-misspelled words with suggestions from Hunspell, selecting the closest correct word. The proposed PAOCR system achieved an impressive accuracy rate of 82.4% on a dataset of printed Arabic characters.

2.1 | Convolutional Neural Networks (CNN) hybrid with other algorithms

In [25], the authors introduce a novel model designed for both single-font and multi-font Arabic text recognition, utilizing Support Vector Machine (SVM) and Convolutional Neural Network (CNN) classifiers. To mitigate overfitting, the model incorporates dropout techniques and performs automatic classification and feature extraction. The authors propose a depth neural network training rule that combines max-margin minimum classification error (M3CE) with cross-entropy methods to enhance performance. The model is evaluated across several databases and compared to state-of-the-art Arabic text recognition methods, demonstrating favorable outcomes and reinforcing its effectiveness in character recognition tasks.

In [26], the researchers explore the application of deep learning and genetic algorithms for recognizing handwritten Arabic characters, addressing the automation of Arabic content recognition, which remains less developed compared to Latin and Chinese content recognition. The study primarily focuses on text processing, with an emphasis on text segmentation and recognition, while discussing the challenges associated with text segmentation and proposing targeted solutions for each challenge. For the recognition phase, a convolutional neural network (CNN) is employed to enhance classification algorithms by automatically extracting features from images. The study evaluates 14 different CNN architectures, achieving a maximum testing accuracy of 91.96% on handwritten Arabic characters. To optimize training parameters, the authors introduce a transfer learning and genetic algorithm approach termed "HMB-AHCR-DLGA," which attains a testing accuracy of 92.88% after conducting five optimization experiments.

In [27], the researchers compare the performance of Bayesian networks and convolutional neural networks (CNNs) for recognizing Arabic handwritten words. The study reveals that the CNN-based system significantly outperforms the Bayesian-based system, achieving an accuracy rate of 96.8% compared to 91.5%. The paper outlines a proposed system that integrates probabilistic graphical models (PGM) with CNN models and details the experiments conducted to assess the performance of these models on the IFN-ENIT database. Overall, this research provides valuable insights into the effectiveness of various machine learning approaches for Arabic handwriting recognition, highlighting the superiority of CNNs in this context.

2.2 Long Short-Term Memory (LSTM)

In [28], the authors compare the performance of 1D and 2D Long Short-Term Memory (LSTM) architectures for recognizing handwritten Arabic characters. They demonstrate that a simple pre-processing step to normalize the position and baseline of the letters allows the 1D LSTM to achieve superior performance while being faster in learning and convergence. The proposed pipeline achieves an accuracy of 91.5% on the IFN/ENIT database, surpassing both manually crafted features with 1D LSTM (87.4%) and 2D LSTM networks (72.9%). Additionally, the authors compare their results with previous studies, which suggested that manually crafted features outperformed automatically learned features (using 1D LSTM) for Arabic handwriting recognition. Their findings, however, showcase the superior performance of their proposed system over manually crafted features in this task.

In [13], the authors introduce a deep learning-based system for recognizing Arabic script, which has been benchmarked on the KHATT dataset. The system is built upon an MDLSTM architecture integrated with a Connectionist Temporal Classification (CTC) layer for alignment. To enhance the input feature space, data augmentation techniques are employed. The model achieves an accuracy of 80% on the KHATT dataset, marking a significant improvement over previous method. The authors attribute this enhancement to the implementation of deep learning and the application of data augmentation techniques.

In [29], the authors propose a novel approach that employs a Long Short-Term Memory (LSTM) network in conjunction with an Elephant Herding Optimization (EHO) algorithm. This method utilizes hybrid feature descriptors, such as Enhanced Local Binary Patterns (ELBP) and Improved Discriminative Model Networks (IDMN), to extract feature values from segmented individual characters. The EHO algorithm is applied to optimize the dimensions of these features, which helps to mitigate overfitting issues and enhances the training and testing processes of the classifier. The optimized features are subsequently inputted into the LSTM network for character classification. The simulation results indicate that the proposed EHO-LSTM model achieves high accuracy rates of 96.66%, 96.67%, and 99.93% for English, Kannada, and Arabic character recognition, respectively, on the chars74K and MADbase digits datasets.

2.3 Other Techniques

In [30], the researchers present a model for Arabic handwriting recognition that integrates the ResNet50 architecture with either Support Vector Machine (SVM) or Random Forest (RF) algorithms. The study reveals that the combination of ResNet50 with Random Forest yields more accurate and consistent results compared to using ResNet50 in isolation. The experimental work encompasses three datasets: the Arabic Handwritten Character Dataset (AHCD), the Alexa Isolated Alphabet Dataset (AIA9K), and the Hijja Dataset. The modified ResNet50 architecture achieves recognition rates of 92.37%, 98.39%, and 91.64% for the three datasets. In contrast, the combined architecture demonstrates improved recognition rates of 95%, 99%, and 92.4% for the same datasets, respectively.

In [31], the study proposes the application of Residual Neural Networks (ResNets) for recognizing Arabic offline isolated handwritten characters. The methodology includes a sequence of steps: pre-processing the data, training the ResNet model, and testing it on various datasets. The approach demonstrates high accuracy levels across three datasets—MADBase, AIA9K, and AHCD—and achieves a notable validation accuracy on a combined dataset. This highlights the effectiveness of ResNets in the domain of Arabic handwriting recognition.

In [32], the authors introduce a comprehensive database of handwritten Arabic-Maghrebi characters and propose a novel approach for recognizing these characters using a deep auto-encoder scheme. The database encompasses all basic Arabic alphabets written in the Maghrebi style, which has not been extensively studied in prior research. The proposed approach achieves a recognition rate of 88% when utilizing a backpropagation neural network with the CENPRMI dataset and 76.54% when employing a combination of multiple Hidden Markov Models (HMMs) classifiers with the IFN/ENIT dataset. This work highlights the significance of developing specialized resources for Arabic-Maghrebi handwriting recognition.

The authors of [33] propose a two-step method for detecting and recovering out-of-vocabulary words in Arabic handwritten text recognition. They compare the effectiveness of this approach to one-step methods that utilize a large static lexicon or a combination of sub-word modeling techniques. The results demonstrate that the proposed two-step approach outperforms the alternatives, achieving an accuracy rate of 91.5% for out-of-vocabulary word detection and 87.3% for out-of-vocabulary word recovery. This research offers valuable insights into enhancing Arabic handwriting recognition systems by effectively addressing the challenges associated with out-of-vocabulary words.

Table 1. Summary of some deep learning applied on Arabic language.

Reference	Feature Extractor	Classifier	Dataset	Size	Accuracy
El-Sawy et al., 2017[34]	CNN	SoftMax	AHCD	16,800	94.9%
Younis, 2017[35]	CNN	SoftMax	AHCD	16,800	97.6%
de Sousa, 2018[36]	CNN	SoftMax	AHCD	16,800	98.42%
Boufenar et al., 2018[37]	CNN	SoftMax	AHCD	16,800	99.98%
Alyahya et al., 2020[38]	CNN	SoftMax	AHCD	16,800	98.3%
Alkhateeb, 2020[39]	CNN	SoftMax	Hijja AHCD	47,434 16,800	92.5% 95.4%
AlJarrah et al., 2021[40]	CNN	SoftMax	AHCD	16,800	97.7%
Altwaijry & Al-Turaiki, 2021a[20]	CNN	SoftMax	Hijja AHCD	47,434 16,800	88% 97%
Alrobah & Albahli, 2021[41]	CNN	SoftMax SVM XGBoost	Hijja	47,434	89% 96.3% 95.7%
Ullah & Jamjoom, 2022[42]	CNN	SoftMax	AHCD	16,800	96.78%
Ali & Mallaiah, 2022[25]	CNN	SVM	AHCD HACDB	16,800 6600	99.71% 99.85%
Nayef et al., 2022[22]	CNN	SoftMax	AHCD Hijja	16,800 47,434	99% 90%
Wagaa et al., 2022[43]	CNN	SoftMax	AHCD Hijja	16,800 47,434	98.48% 95%
Bouchriha et al., 2022[44]	CNN	SoftMax	Hijja	47,434	95%
Bin Durayhim et al. 2023 [45]	CNN	SoftMax	AHCD Hijja	16,800 47,434	98% 99.5%

2.4 Arabic Datasets

2.4.1 Arabic Handwritten Characters Data set (AHCD)

The Arabic Handwritten Characters Dataset1 (AHCD) [34] comprises a vast collection of handwritten Arabic script characters, encompassing a wide variety of both standalone characters and those embedded within words. This dataset effectively captures the diverse writing styles and nuances of Arabic handwriting, containing a total of 16,800 characters produced by 60 individuals aged between 19 and 40, with 90% of participants being right-handed. Each character, ranging from 'alef' to 'yeh', was written by each participant ten times in two different formats, with the resulting forms scanned at a resolution of 300 dpi.

The dataset is organized into two subsets: a training set consisting of 13,440 characters (480 images per class) and a test set comprising 3,360 characters (120 images per class). Importantly, there is no overlap between the writers in the training and test sets; the selection of writers for the test set was randomized to ensure variability and avoid bias from a single institution.

The creation of this dataset encountered several challenges, including variations in writing style, thickness, the number and placement of dots, and inconsistencies in character shapes when written in similar positions. To mitigate these challenges, various pre-processing techniques were applied to

[1] https://www.kaggle.com/datasets/mloey1/ahcd1

reduce noise and enhance the legibility of the input images. The pre-processing stage is vital for any recognition system and typically includes converting RGB images to grayscale, as well as employing filtering and smoothing techniques.

2.4.2 Hijjaa Data set

The Hijjaa dataset1 is a publicly accessible collection of individual Arabic letters recently introduced by Altwaijry et al. [20]. This dataset was created by Saudi Arabian schoolchildren aged between seven and twelve in Riyadh and consists of 108 classes, with each class representing a distinct Arabic letter. The letters are displayed in four different contexts: isolated, at the beginning, in the middle, and at the end of a word. In total, the Hijjaa dataset comprises 47,434 images.

Data collection from children presented numerous challenges. Participants often had difficulty following the reference materials, resulting in missing characters, incorrectly placed letters, and repeated entries. Common issues included visible marks and erased pencil strokes, which required manual cleaning, while some faint pencil strokes needed enhancement for clarity. Additionally, some scanned pages were tilted and had to be manually aligned for proper orientation.

The initial scanned raw PNG images were divided into 108 square PNG images, each containing a single letter form sized at 256x256 pixels. These images were subsequently resized to 32x32 pixels using Python. However, due to the partial filling of many matrices, the dataset exhibited an imbalanced representation of letters, with a higher number of PNG files available for the first half of the alphabet compared to the latter. Although the expected number of images from the 591 scanned matrices was 63,828, unfilled matrix cells were discarded, leading to the random selection of a subset of PNG files from each letter to create a more balanced dataset. Ultimately, the dataset contains 47,434 characters.

Each Arabic letter is organized into 29 folders, representing different classes, with a dedicated folder for the Arabic letter "hamza." Within each class folder, subfolders are created for the various letter forms, housing the images corresponding to each specific letter form. It is important to note that the dataset does not include vocalization diacritics (harakat), which are used to indicate vowels and other phonetic sounds not represented by the Arabic letters.

2.5 Other Languages

Additionally, languages such as Farsi, Kurdish, Urdu, and Pashto incorporate words, features, and structural elements derived from Arabic.

2.5.1 Urdu

Urdu, recognized as the fifth most spoken language globally, accounts for 4.7 percent of the world's population and is predominantly spoken in Pakistan as the national language, as well as in India, where it is one of the 22 official languages. Urdu speakers are found in 20 countries, including India, Pakistan, Turkey, Saudi Arabia, Bangladesh, Afghanistan, Iran, Azerbaijan, and Nepal. The Urdu language is a synthesis of multiple languages, including Arabic, Persian, Turkish, and Hindi, and features a non-scripting nature.

Due to its cursive style and significant influence from Arabic and Persian scripts, Urdu does not differentiate between capital and lowercase letters, resulting in writing similarities across these languages. Urdu is classified as a bidirectional language because it employs two writing systems: the Urdu script, which is written from right to left, and numerals, which are written from left to right. The Urdu script comprises 38 basic letters and 10 numeric letters. This character set also encompasses other Urdu-based scripts, such as Arabic (which contains 28 characters) and Persian (which includes 32 characters).

[1] https://github.com/israksu/Hijja2

Each character in Urdu can take on multiple forms—isolated, initial, medial, and final—depending on its position within a word. For example, the characters Beh (ب), Peh (پ), and Teh (ت) can connect with neighboring characters from either side. In contrast, characters such as Alif (ا), Daal (د), and Reh (ر) only connect from right to left, with preceding characters in a word. Additionally, some characters, like Hamza (ء), lack any joining ability. Based on these joining properties, Urdu characters can be categorized into two groups: joiners, which can adopt all four forms depending on their neighboring characters, and non-joiners, which can only take the isolated and final shapes.

In [46], a deep neural network-based system is introduced for recognizing handwritten Urdu characters, trained on a dataset comprising 74,285 samples and tested on 21,223 samples. The system achieved an impressive recognition rate of 98.82% across 133 classes, outperforming existing state-of-the-art systems for Urdu recognition. Furthermore, it was tested on datasets containing numeral characters from five different languages, resulting in an average recognition accuracy of 99.26% with a precision of 99.29%, and a recognition accuracy of 99.322% for each individual language.

In [47], researchers present a hybrid approach that combines convolutional neural networks (CNN) and multi-dimensional long short-term memory networks (MDLSTM) for recognizing cursive Urdu Nastaliq script. The CNN extracts low-level features, which are then processed by MDLSTM for contextual learning. The methodology was tested on the publicly available Urdu Printed Text-line Image (UPTI) dataset, achieving an accuracy of 98.12% for 44 classes, thus surpassing state-of-the-art results on the UPTI dataset.

In [48], the authors propose a hybrid deep learning approach utilizing an encoder-decoder structure. The approach includes a CNN for feature extraction, a bi-directional Gated Recurrent Unit (BiGRU) as the encoder, and a Gated Recurrent Unit (GRU) as the decoder to recognize printed Urdu script in Nastaleeq font. The dataset was split into three parts: 50% for training, 30% for validation, and 20% for testing. The model achieved an accuracy of 98.5% on the test dataset across two types of experiments, one considering variations based on character position, resulting in 191 unique categories, and the other focusing solely on 99 basic categories.

In [49], a Conv-transformer architecture is proposed for recognizing unconstrained offline Urdu handwriting. The model first employs a CNN, followed by a vanilla transformer. The convolution layers reduce spatial resolutions and address complexities associated with transformer multi-head attention layers. The model, trained on both printed and handwritten Urdu text lines, achieved a character error rate (CER) of 5%. However, the authors highlight that the limited amount of training data poses a significant challenge, suggesting that a transformer without convolution layers might be viable with more data.

In [50], the authors introduce an implicit segmentation approach based on multi-dimensional long short-term memory (MDLSTM) networks for recognizing Urdu Nastaliq text. The proposed system outperforms state-of-the-art Urdu text line recognition systems, achieving a recognition accuracy of 98±0.25%.

In [51], researchers present a recognition system utilizing implicit segmentation for Urdu text lines in Nastaliq script. The system employs a multi-dimensional long short-term memory recurrent neural network (MDLSTM RNN) with a connectionist temporal classification (CTC) output layer. The approach involves sliding overlapped windows over text lines to extract statistical features, achieving a promising recognition rate of 94.5% on the UPTI database.

In [52], a multi-dimensional recurrent neural network and statistical features are utilized to recognize Urdu Nastaliq text through a three-stage process: pre-processing and feature extraction, MDLSTM processing, and CTC output. The system, tested on the UPTI dataset, achieved an accuracy of 91.5%. However, the authors acknowledge limitations, including the system's inability to account for shape variations affecting recognition accuracy, reliance on a limited dataset representative only of printed Urdu text images, and lack of comparisons with other state-of-the-art systems for Urdu Nastaliq text recognition.

2.5.2 Pashto

Pashto is an Indo-Iranian language that is classified under the larger Indo-European language family. It features a complex system of dialects, the two primary groups being the Northern and Southern dialects. Notable variants include Kandahari, prevalent in the south, and Peshawari, spoken in the northern regions. These dialects exhibit distinctive phonetic and grammatical characteristics, which contribute to the language's rich linguistic diversity.

The geographic distribution of Pashto is primarily concentrated in Afghanistan and Pakistan. In Afghanistan, Pashto holds the status of one of the two official languages, alongside Dari (Persian). Within Pakistan, it is predominantly spoken in the Khyber Pakhtunkhwa province and parts of Balochistan. Estimates indicate that the number of native Pashto speakers ranges from approximately 50 to 60 million, reflecting its significance as a major regional language.

Pashto employs a modified Arabic script, which includes additional letters specifically designed to accommodate the unique phonetic sounds of the language. The script's adaptations facilitate accurate representation of Pashto phonology. Moreover, Pashto possesses a rich literary tradition, particularly in poetry, with historical figures such as Khushal Khan Khattak and Gul Khan Nasir making significant contributions to its literary canon.

Beyond its linguistic attributes, Pashto serves as a crucial marker of ethnic identity for the Pashtun people, who are characterized by their distinct cultural heritage and traditional practices. The language is interwoven with the community's folklore and music, reflecting central themes such as love, valor, and honor. This cultural richness underscores the importance of Pashto in maintaining the social fabric of Pashtun society.

In contemporary contexts, Pashto is utilized across various media platforms, including television, radio, and digital outlets, with dedicated Pashto language television channels enhancing its accessibility. Furthermore, language learning resources, such as textbooks and online courses, are increasingly available, facilitating the acquisition of Pashto for both native and non-native speakers.

Despite its prominence, Pashto faces challenges, particularly in terms of standardization. The existence of diverse dialects complicates the establishment of a unified standard for educational and official purposes. Additionally, the socio-political dynamics of Afghanistan and Pakistan significantly influence the language's status, impacting its development, preservation, and the overall vitality of Pashto in the face of modern challenges.

3 | Conclusion

The study explores the landscape of Arabic Optical Character Recognition (OCR), emphasizing its significance for Arab and Muslim communities. Despite the advancements in OCR technology, most current systems operate offline and struggle with real-time recognition of Arabic text due to inherent complexities in the language and character structure. Many existing approaches are confined to private datasets or are limited to recognizing words and paragraphs, hindering a comprehensive assessment of their real-world performance.

The main objectives of the study include:

- Critical Analysis: A thorough review of existing research on Arabic OCR was conducted to identify major trends, issues, and challenges faced in the field. The analysis revealed that many studies lack a sequential approach to performance improvement and do not build on previous research, leading to unreliable results.

- Evaluation of Deep Learning Techniques: The study systematically evaluates deep learning techniques for feature extraction, particularly focusing on recognition performance. It was found that hybrid feature extraction and classification methods could significantly enhance Arabic OCR accuracy.

- Future Directions: The research proposes that future work should focus on customizing and training established Convolutional Neural Networks (CNNs) such as GoogLeNet, VGGNet, AlexNet, and DenseNet. Training these models on large datasets can improve recognition rates. Additionally, combining CNNs with Recurrent Neural Networks (RNNs) is suggested to facilitate the recognition of both handwritten and printed Arabic text.
- Standard Benchmark Evaluation: The study emphasizes the importance of evaluating proposed OCR approaches on standard benchmarks to measure real-world performance effectively.

The findings of this study highlight the necessity of hybrid approaches and comprehensive evaluations in advancing Arabic OCR technology. The study also notes the limitations of previous research, including a lack of temporal order and foundational building on past work, leading to unreliable results. The authors plan to apply and compare some of these studies, providing access to scientific codes to validate their findings in future research.

Acknowledgments

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Alghyaline, S. (2022). Arabic Optical Character Recognition: A Review. Computer Modeling in Engineering & Sciences, 135(3), 1825–1861. https://doi.org/10.32604/CMES.2022.024555

[2] Al-Ayyoub, M., Nuseir, A., Alsmearat, K., Jararweh, Y., & Gupta, B. (2018). Deep learning for Arabic NLP: A survey. Journal of Computational Science, 26, 522–531. https://doi.org/10.1016/J.JOCS.2017.11.011

[3] IBM Cloud Education. (2022, January 5). What Is Optical Character Recognition (OCR)? | IBM. IBM. https://www.ibm.com/cloud/blog/optical-character-recognition

[4] Altwaijry, N., & Al-Turaiki, I. (2021). Arabic handwriting recognition system using convolutional neural network. Neural Computing and Applications, 33(7), 2249–2261. https://doi.org/10.1007/S00521-020-05070-8

[5] Aljarrah, M. N., Zyout, M. M., & Duwairi, R. (2021). Arabic Handwritten Characters Recognition Using Convolutional Neural Network. 2021 12th International Conference on Information and Communication Systems, ICICS 2021, 182–188. https://doi.org/10.1109/ICICS52457.2021.9464596

[6] Aljarrah, M. N., Zyout, M. M., & Duwairi, R. (2021). Arabic Handwritten Characters Recognition Using Convolutional Neural Network. 2021 12th International Conference on Information and Communication Systems, ICICS 2021, 182–188. https://doi.org/10.1109/ICICS52457.2021.9464596

[7] Wu, X., Hong, D., & Chanussot, J. (2023). UIU-Net: U-Net in U-Net for Infrared Small Object Detection. IEEE Transactions on Image Processing, 32, 364–376. https://doi.org/10.1109/TIP.2022.3228497

[8] Hong, D., Gao, L., Yao, J., Zhang, B., Plaza, A., & Chanussot, J. (2021). Graph Convolutional Networks for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 59(7), 5966–5978. https://doi.org/10.1109/TGRS.2020.3015157

[9] Wick, C., Reul, C., & Puppe, F. (2018). Calamari-a high-performance tensorflow-based deep learning package for optical character recognition. ArXiv Preprint ArXiv:1807.02004.

[10] O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., Riordan, D., & Walsh, J. (2020). Deep Learning vs. Traditional Computer Vision. Advances in Intelligent Systems and Computing, 943, 128–144. https://doi.org/10.1007/978-3-030-17795-9_10

[11] Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature 2015 521:7553, 521(7553), 436–444. https://doi.org/10.1038/nature14539

[12] Malik, S., Sajid, A., Ahmad, A., Almogren, A., Hayat, B., Awais, M., & Kim, K. H. (2020). An efficient skewed line segmentation technique for cursive script OCR. Scientific Programming, 2020. https://doi.org/10.1155/2020/8866041

[13] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep Learning for Generic Object Detection: A Survey. International Journal of Computer Vision, 128(2), 261–318. https://doi.org/10.1007/S11263-019-01247-4

[14] Abdalkafor, A. S. (2018). Survey for Databases on Arabic Off-line Handwritten Characters Recognition System. 1st International Conference on Computer Applications and Information Security, ICCAIS 2018. https://doi.org/10.1109/CAIS.2018.8442001

[15] Soumia, F., Djamel, G., & Haddad, M. (2021). Handwritten Arabic Character Recognition: Comparison of Conventional Machine Learning and Deep Learning Approaches. Lecture Notes on Data Engineering and Communications Technologies, 72, 1127–1138. https://doi.org/10.1007/978-3-030-70713-2_100

[16] Owens, J. (2013). The Oxford handbook of Arabic linguistics. Oxford University Press.

[17] Ahmed, R., Dashtipour, K., Gogate, M., Raza, A., Zhang, R., Huang, K., Hawalah, A., Adeel, A., & Hussain, A. (2020). Offline arabic handwriting recognition using deep machine learning: A review of recent advances. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11691 LNAI, 457–468. https://doi.org/10.1007/978-3-030-39431-8_44

[18] Alrobah, N., & Albahli, S. (2022). Arabic Handwritten Recognition Using Deep Learning: A Survey. Arabian Journal for Science and Engineering, 47(8), 9943–9963. https://doi.org/10.1007/s13369-021-06363-3

[19] Alghyaline, S. (2023). Arabic Optical Character Recognition: A Review. In CMES - Computer Modeling in Engineering and Sciences (Vol. 135, Issue 3, pp. 1825–1861). Tech Science Press. https://doi.org/10.32604/cmes.2022.024555

[20] Altwaijry, N., & Al-Turaiki, I. (2021). Arabic handwriting recognition system using convolutional neural network. Neural Computing and Applications, 33(7), 2249–2261. https://doi.org/10.1007/s00521-020-05070-8

[21] Balaha, H. M., Ali, H. A., Saraya, M., & Badawy, M. (2021). A new Arabic handwritten character recognition deep learning system (AHCR-DLS). Neural Computing and Applications, 33(11), 6325–6367. https://doi.org/10.1007/s00521-020-05397-2

[22] Nayef, B. H., Abdullah, S. N. H. S., Sulaiman, R., & Alyasseri, Z. A. A. (2022). Optimized leaky ReLU for handwritten Arabic character recognition using convolution neural networks. Multimedia Tools and Applications, 81(2), 2065–2094. https://doi.org/10.1007/s11042-021-11593-6

[23] Elkhayati, M., & Elkettani, Y. (2022). UnCNN: A New Directed CNN Model for Isolated Arabic Handwritten Characters Recognition. Arabian Journal for Science and Engineering, 47(8), 10667–10688. https://doi.org/10.1007/s13369-022-06652-5

[24] Alghyaline, S. (2022). A Printed Arabic Optical Character Recognition System using Deep Learning. Journal of Computer Science, 18(11), 1038–1050. https://doi.org/10.3844/jcssp.2022.1038.1050

[25] Ali, A. A. A., & Mallaiah, S. (2022). Intelligent handwritten recognition using hybrid CNN architectures based-SVM classifier with dropout. Journal of King Saud University - Computer and Information Sciences, 34(6), 3294–3300. https://doi.org/10.1016/J.JKSUCI.2021.01.012

[26] Balaha, H. M., Ali, H. A., Youssef, E. K., Elsayed, A. E., Samak, R. A., Abdelhaleem, M. S., Tolba, M. M., Shehata, M. R., Mahmoud, M. R., Abdelhameed, M. M., & Mohammed, M. M. (2021). Recognizing arabic handwritten characters using deep learning and genetic algorithms. Multimedia Tools and Applications, 80(21–23), 32473–32509. https://doi.org/10.1007/s11042-021-11185-4

[27] Khémiri, A., Echi, A. K., & Elloumi, M. (2019). Bayesian Versus Convolutional Networks for Arabic Handwriting Recognition. Arabian Journal for Science and Engineering, 44(11), 9301–9319. https://doi.org/10.1007/s13369-019-03939-y

[28] Yousefi, M. R., Soheili, M. R., Breuel, T. M., & Stricker, D. (2015). A comparison of 1D and 2D LSTM architectures for the recognition of handwritten Arabic. Document Recognition and Retrieval XXII, 9402, 94020H. https://doi.org/10.1117/12.2075930

[29] Guptha, N. S., Balamurugan, V., Megharaj, G., Sattar, K. N. A., & Rose, J. D. (2022). Cross lingual handwritten character recognition using long short term memory network with aid of elephant herding optimization algorithm. Pattern Recognition Letters, 159, 16–22. https://doi.org/10.1016/j.patrec.2022.04.038

[30] Khudeyer, R. S., & Al-Moosawi, N. M. (2023). Combination of machine learning algorithms and Resnet50 for Arabic Handwritten Classification. Informatica, 46(9), 39–44. https://doi.org/10.31449/INF.V46I9.4375

[31] Al-Taani, A. T., & Ahmad, S. T. (2021). RECOGNITION OF ARABIC HANDWRITTEN CHARACTERS USING RESIDUAL NEURAL NETWORKS. Jordanian Journal of Computers and Information Technology (JJCIT), 07(02).

[32] Djaghbellou, S., Attia, A., Bouziane, A., & Akhtar, Z. (2022). Local features enhancement using deep auto-encoder scheme for the recognition of the proposed handwritten Arabic-Maghrebi characters database. Multimedia Tools and Applications, 81(22), 31553–31571. https://doi.org/10.1007/s11042-022-13032-6

[33] Jemni, S. K., Kessentini, Y., & Kanoun, S. (2019). Out of vocabulary word detection and recovery in Arabic handwritten text recognition. Pattern Recognition, 93, 507–520. https://doi.org/10.1016/j.patcog.2019.05.003

[34] El-Sawy, A., Loey, M., & El-Bakry, H. (2017). Arabic handwritten characters recognition using convolutional neural network. WSEAS Transactions on Computer Research, 5(1), 11–19.

[35] Younis, K. S. (2017). Arabic hand-written character recognition based on deep convolutional neural networks. Jordanian Journal of Computers and Information Technology, 3(3).

[36] de Sousa, I. P. (2018). Convolutional ensembles for Arabic Handwritten Character and Digit Recognition. PeerJ Computer Science, 2018(10), e167. https://doi.org/10.7717/PEERJ-CS.167/SUPP-4

[37] Boufenar, C., Kerboua, A., & Batouche, M. (2018). Investigation on deep learning for off-line handwritten Arabic character recognition. Cognitive Systems Research, 50, 180–195. https://doi.org/10.1016/J.COGSYS.2017.11.002

[38] Alyahya, H., Ismail, M. M. Ben, & Al-Salman, A. (2020). Deep ensemble neural networks for recognizing isolated Arabic handwritten characters. ACCENTS Transactions on Image Processing and Computer Vision, 6(21), 68–79. https://doi.org/10.19101/TIPCV.2020.618051

[39] Alkhateeb, J. H. (2020). An effective deep learning approach for improving off-line arabic handwritten character recognition. International Journal of Software Engineering and Computer Systems, 6(2), 53–61.

[40] AlJarrah, M. N., Mo'ath, M. Z., & Duwairi, R. (2021). Arabic handwritten characters recognition using convolutional neural network. 2021 12th International Conference on Information and Communication Systems (ICICS), 182–188.

[41] Alrobah, N., & Albahli, S. (2021). A Hybrid Deep Model for Recognizing Arabic Handwritten Characters. IEEE Access, 9, 87058–87069. https://doi.org/10.1109/ACCESS.2021.3087647

[42] Ullah, Z., & Jamjoom, M. (2022). An intelligent approach for Arabic handwritten letter recognition using convolutional neural network. PeerJ Computer Science, 8, e995. https://doi.org/10.7717/PEERJ-CS.995/SUPP-2

[43] Wagaa, N., Kallel, H., & Mellouli, N. (2022). Improved Arabic Alphabet Characters Classification Using Convolutional Neural Networks (CNN). Computational Intelligence and Neuroscience, 2022, 9965426. https://doi.org/10.1155/2022/9965426

[44] Bouchriha, L., Zrigui, A., Mansouri, S., Berchech, S., & Omrani, S. (2022). Arabic Handwritten Character Recognition Based on Convolution Neural Networks. Communications in Computer and Information Science, 1653 CCIS, 286–293. https://doi.org/10.1007/978-3-031-16210-7_23

[45] Bin Durayhim, A., Al-Ajlan, A., Al-Turaiki, I., & Altwaijry, N. (2023). Towards Accurate Children's Arabic Handwriting Recognition via Deep Learning. Applied Sciences 2023, Vol. 13, Page 1692, 13(3), 1692. https://doi.org/10.3390/APP13031692

[46] Mushtaq, F., Misgar, M. M., Kumar, M., & Khurana, S. S. (2021). UrduDeepNet: offline handwritten Urdu character recognition using deep neural network. Neural Computing and Applications, 33(22), 15229–15252. https://doi.org/10.1007/S00521-021-06144-X

[47] Naz, S., Umar, A. I., Ahmad, R., Siddiqi, I., Ahmed, S. B., Razzak, M. I., & Shafait, F. (2017). Urdu Nastaliq recognition using convolutional–recursive deep learning. Neurocomputing, 243, 80–87. https://doi.org/10.1016/J.NEUCOM.2017.02.081

[48] Zia, S., Azhar, M., Lee, B., Tahir, A., Ferzund, J., Murtaza, F., & Ali, M. (2023). Recognition of printed Urdu script in Nastaleeq font by using CNN-BiGRU-GRU Based Encoder-Decoder Framework. Intelligent Systems with Applications, 18, 200194. https://doi.org/10.1016/J.ISWA.2023.200194

[49] Riaz, N., Arbab, H., Maqsood, A., Nasir, K., Ul-Hasan, A., & Shafait, F. (2022). Conv-transformer architecture for unconstrained off-line Urdu handwriting recognition. International Journal on Document Analysis and Recognition, 25(4), 373–384. https://doi.org/10.1007/S10032-022-00416-5

[50] Naz, S., Umar, A. I., Ahmed, R., Razzak, M. I., Rashid, S. F., & Shafait, F. (2016). Urdu Nasta'liq text recognition using implicit segmentation based on multi-dimensional long short term memory neural networks. SpringerPlus, 5(1), 1–16. https://doi.org/10.1186/S40064-016-3442-4

[51] Naz, S., Umar, A. I., Ahmad, R., Ahmed, S. B., Shirazi, S. H., Siddiqi, I., & Razzak, M. I. (2016). Offline cursive Urdu-Nastaliq script recognition using multidimensional recurrent neural networks. Neurocomputing, 177, 228–241. https://doi.org/10.1016/J.NEUCOM.2015.11.030

[52] Naz, S., Umar, A. I., Ahmad, R., Ahmed, S. B., Shirazi, S. H., & Razzak, M. I. (2017). Urdu Nasta'liq text recognition system based on multi-dimensional recurrent neural network and statistical features. Neural Computing and Applications, 28(2), 219–231. https://doi.org/10.1007/S00521-015-2051-4