HICMA: The Handwriting Identification for Calligraphy and Manuscripts in Arabic Dataset

Anis Ismail
KU Leuven
Leuven, Belgium
anis.ismail@student.kuleuven.be

Zina Kamel
Lebanese American University
Beirut, Lebanon
zina.kamel@lau.edu

Reem Mahmoud
intervu.ai
Richardson, TX, USA
reem.mahmoud@intervu.ai

Abstract

Arabic is one of the most globally spoken languages with more than 313 million speakers worldwide. Arabic handwriting is known for its cursive nature and the variety of writing styles used. Despite the increase in effort to digitize artistic and historical elements, no public dataset was released to deal with Arabic text recognition for realistic manuscripts and calligraphic text. We present the Handwriting Identification of Manuscripts and Calligraphy in Arabic (HICMA) dataset as the first publicly available dataset with real-world and diverse samples of Arabic handwritten text in manuscripts and calligraphy. With more than 5,000 images across five different styles, the HICMA dataset includes image-text pairs and style labels for all images. We further present a comparison of the current state-of-the-art optical character recognition models in Arabic and benchmark their performance on the HICMA dataset, which serves as a baseline for future works. Both the HICMA dataset and its benchmarking tool are made available to the public under the CC BY-NC 4.0 license in the hope that the presented work opens the door to further enhancements of complex Arabic text recognition.

1 Introduction

Handwriting is a method used by humans to convey information in a written medium. Every person possesses a unique style when drawing characters. This leads to a wide variation in the expression of written characters and texts. Arabic text is of particular interest as Arabic is one of the most globally spoken languages with more than 313 million speakers worldwide. In the Arabic language, the complexity of written text increases since each character inherently has different forms depending on its position in the word, that is, whether it is in the beginning, middle, or end of the word.

Historical Arabic text is abundant with more than ten centuries of rich Arabic history and is often in need of being digitized. Arabic historical manuscripts typically encompass handwritten texts, often of a significant age, characterized by cursive script, varying styles, and various artistic intricacies surrounding the written text. Arabic calligraphy is a special form of Arabic handwriting often used in manuscripts and as a prominent tool for ornating architecture. The Arabic language relies on a variety of styles in manuscripts and calligraphy, each providing a different level of aesthetic artistic views and possessing its own rules. The most popular styles of handwriting in Arabic manuscripts and calligraphy are Diwani, Thuluth, Kufic, Farsi, Naskh, and Ruqaa. Arabic calligraphy is usually hand-drawn by experienced artists with complex drawing techniques that include heavy use of diacritics and decorative symbols. Consequently, non-expert readers struggle to understand the calligraphic text.

Handwriting recognition is the task involved in converting handwritten text, which is typically captured as images, into machine-readable text. The complexity of this task is in accurately recognizing variations in the different styles of writing. Moreover, the complexity becomes more apparent in historical Arabic handwritten text due to its nature. To address the challenges in handwritten Arabic and enhance the accessibility of Arabic calligraphic content, the development of models capable of accurately recognizing this intricate handwritten text becomes essential. This, in turn, necessitates the availability of large datasets for the training and validation of such models. Many works focused on creating datasets for the task of style classification of Arabic calligraphy, such as the work of Kaoudja et al.'s (2019), while others focused on creating datasets for single character recognition (Altwaijry and Al-Turaiki, 2021), Alrehali et al.'s (2020)[1] or single-digit recognition (Abdelazeem

[1] The dataset is a combination of 3 subsets containing each 2,240, 1,000 and 2,000 characters

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 24–32
December 7, 2023 ©2023 Association for Computational Linguistics

and El-Sherif, 2017). The Calliar dataset (Alyafeai et al., 2021) is the only existing dataset today that is tailored for Arabic calligraphy recognition, on the character, word, sentence, and stroke levels. This dataset, however, contains calligraphic text drawn using digital pens on a plain white background, eliminating the realistic calligraphy style found in real-world Arabic scripts.

Despite the plethora of datasets available in the Arabic handwriting recognition space, very few represent a realistic and rich variety of styles for both historic manuscripts and calligraphy, target full-sentence handwriting recognition from unprocessed images, and are publicly accessible. We present the first publicly available dataset for Arabic handwritten text in both manuscripts and calligraphy forms called the Handwriting Identification for Calligraphy and Manuscripts in Arabic (HICMA) Dataset. With more than 5,000 images across five different Arabic writing styles, the HICMA dataset includes image-text pairs and style labels for all images. In this manuscript, we describe the collection, labeling, and processing steps of the novel HICMA dataset and present a benchmark evaluation of the latest Optical Character Recognition (OCR) models for the Arabic language on HICMA. The contributions of our work are three-fold:

1. We present the first publicly available Arabic handwriting recognition dataset targeting full sentence recognition from unprocessed images.

2. We introduce an Arabic handwriting recognition dataset that is among the most diverse collections of Arabic historic manuscripts and calligraphy with more than 5,000 images across five different writing styles.

3. We preserve the contextual details and artistic styles of the Arabic manuscripts and calligraphic text in our dataset to closely represent the occurrence of such text in real-world materials.

We make the HICMA dataset[2] and the benchmarking tool[3] presented in this manuscript publicly accessible to the research community.

2 Related Work

Several studies have dealt with collecting various types of datasets for different formats of Arabic handwriting. For regular Arabic handwriting, there are many datasets present in literature such as KHATT (Mahmoud et al., 2018), consisting of 1,000 handwritten forms collected across 1,000 different writers from different countries. It was then extended to the Online-KHATT (Mahmoud et al.) dataset consisting of 10,040 lines of handwritten text by 623 different writers. ADAB (Märgner and El Abed, 2009) is another dataset that consists of 32,492 Arabic words handwritten by more than 1,000 writers. There are also multilingual datasets that combine Arabic and English like MAYASTROUN (Njah et al., 2012), which consists of 67,825 samples written by 355 writers. The MAYASTROUN dataset consists of varying script types including words, characters, digits, mathematical expressions, and signatures.

In contrast to regular Arabic handwriting datasets, few studies in the literature have dealt with Arabic manuscript and calligraphy text. One important dataset for Arabic calligraphy is the Calliar dataset (Alyafeai et al., 2021) which records digitized versions of images as strokes and drawings using digital pens. Calliar is annotated for stroke, character, word, and sentence-level prediction. It also consists of 45,572 strokes, 7,556 words, and 2,500 sentences. However, the resulting dataset overlooks the contextual details present in real-world calligraphy such as the texture of the paper, surrounding artistic styles, noise, and interactions with other elements in the artwork. This as a result impacts an Optical Character Recognition (OCR) model's ability to recognize calligraphy in diverse and authentic settings.

Other datasets in literature targeted calligraphy style classification by focusing on the style classification alone such as the dataset by Kaoudja et al.'s (2019). Kaoudja et al. (2019) collected 1,685 images and classified them into 9 different calligraphic styles including Thuluth, Naskh, and Diwani. Each calligraphy style consists of around 180 to 195 images. Moreover, Allaf and Al-Hmouz (2016) developed a dataset and designed a system for classifying calligraphy images with artistic Arabic calligraphy types, mainly Thuluth, Reqaa, and

[2] https://hicma.net/

[3] https://github.com/anisdismail/HICMA-benchmark

Dataset	Size	Data Type	Number of Styles	Data Public
Alrehali et al.'s (2020)	5,240	characters	1 (Naskh)	×
MADbase (Abdelazeem and El-Sherif, 2017)	70,000	digits	unspecified	✓
KHATT (Mahmoud et al., 2018)	4,000	paragraphs	unspecified	✓
Calliar (Alyafeai et al., 2021)	2,500/40,000	sentences /strokes	4	✓
ADAB (Märgner and El Abed, 2009)	32,492	words	unspecified	✓
Hijja (Altwaijry and Al-Turaiki, 2021)	47,434	characters	unspecified	✓
Kaoudja et al.'s (2019)	1,685	sentences	9	×
Allaf and Al-Hmouz's (2016)	267	sentences	3	✓
KERTAS (Adam et al., 2018)	2,000	letters	unspecified	✓
Salamah and King's (2018)	1,000	letters	10	✓
Khayyat and Elrefaei's (2020)	8,638	pages	unspecified	×
MAYASTROUN (Njah et al., 2012)	67,825	varied	unspecified	×
HICMA (Ours)	5,031	sentences/styles	5	✓

Table 1: Summary of Available Datasets in Literature

Kufi. Their dataset consists of 267 images divided evenly across the three calligraphy types. Salamah and King (2018) also approached the challenge of calligraphy style classification and collected 1,000 calligraphy images scraped from public websites in various calligraphy styles. Other sophisticated datasets, such as KERTAS (Adam et al., 2018), studied images of historical manuscripts. For producing KERTAS, 2,000 images were taken from various handwritten Arabic scripts dating back to the fourteenth century and were manually annotated and segmented to extract images of the characters in the text. Furthermore, Khayyat and Elrefaei (2020) collected 8,638 images of historical Arabic manuscripts. Their dataset is categorized into fourteen classes with six handwriting styles. Adam et al. (2017) collected 330 images of isolated Arabic letters that were extracted from ancient manuscripts. This dataset consists of Ruqaa, Diwani, Kufi, Naskh, and Farsi styles and has been used to classify Arabic script styles based on segmented letters.

The aforementioned calligraphy works can be classified into two categories, (a) datasets that simplified calligraphy for recognition tasks and (b) datasets that focused only on style classification with authentic calligraphy text. The simplified calligraphy datasets removed the contextual details commonly seen in real-world calligraphy. The remaining datasets that preserved the calligraphy in its true form were focused only on style classification, making them not directly useful for handwriting recognition. To the best of our knowledge, there is no dataset in the literature that deals with Arabic handwriting recognition in both manuscript and calligraphy images. Furthermore, many of the aforementioned datasets were either not publicly available or did not allow tampering with their dataset content. This makes the majority of the datasets in the literature not readily accessible for

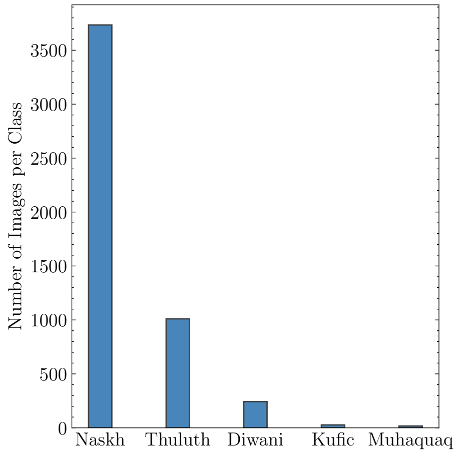


Figure 1: The style distribution of Arabic text across the HICMA dataset.

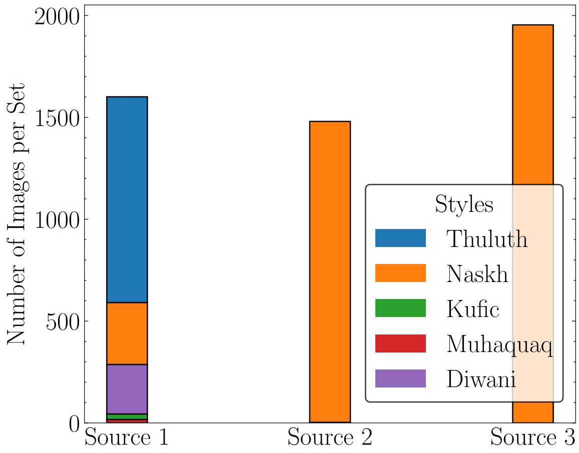


Figure 2: The style distribution of Arabic text per the 3 data sources of HICMA.

research purposes.

In Table 1, we present a comparative analysis of the existing datasets based on five criteria namely size, data type, number of styles, and whether the dataset is publicly available or not. In this paper, we introduce the HICMA dataset that targets both Arabic manuscripts and calligraphy handwriting recognition while preserving the artistic styles and contextual details of the calligraphy to closely represent real-world data.

3 HICMA Dataset

3.1 Data Collection

The first step of creating the HICMA dataset was collecting the images of the handwritten Arabic text. We collected images with various calligraphy styles including Thuluth, Diwani, Muhaquaq, Naskh, and Kufic. We relied on the following resources for building our dataset:

- **Source 1**: The Free Islamic Calligraphy website[4], which represents a Jordanian non-governmental organization (NGO) dedicated to sharing Islamic calligraphy paintings for free in a variety of styles.

- **Source 2**: The Ibn Bawab Qur'an from the Chester Beatty Library[5] located in Dublin, Ireland. This Qur'an is one of the oldest versions of the Qur'an that is written in the Naskh style by Abu'l-Hasan 'Ali ibn Hilal, who was known as Ibn al-Bawwab in the 11th century. We selected 106 pages of the Qur'an text with each page containing around 15 lines.

- **Source 3**: A private collection of manuscripts and religious writings in Naskh style dating back to the 17th century, which were made accessible by courtesy of Dr. Vahid Behmardi. We photographed and collected manuscripts of 202 available pages.

Permission was granted from all the above resources to publish all collected images in a dataset for academic research purposes.

3.2 Data Labeling

For the labeling process, 11 volunteers were recruited and trained to support in reading and recording the Arabic text in the images. The volunteers were divided into two teams who worked on labeling different images in parallel. Both teams started working on source 1, followed by source 2, and finally source 3. Every set was divided among the two teams, and once a team labeled their corresponding subset, the other team would validate the opposing team's labels. This cross-validation technique is employed to improve the quality of the produced labels and ensure accurate labels.

After the labeling process was finished, the images were processed to remove duplicate samples as well as remove diacritics and punctuation using the pyArabic[6] package. The prepared dataset was then divided into training, validation, and testing sets following an 80%-10%-10% division, respectively. To ensure that the three resulting sets have

[4] https://freeislamiccalligraphy.com

[5] https://viewer.cbl.ie/viewer/image/Is_1431/1/

[6] https://pypi.python.org/pypi/pyarabic

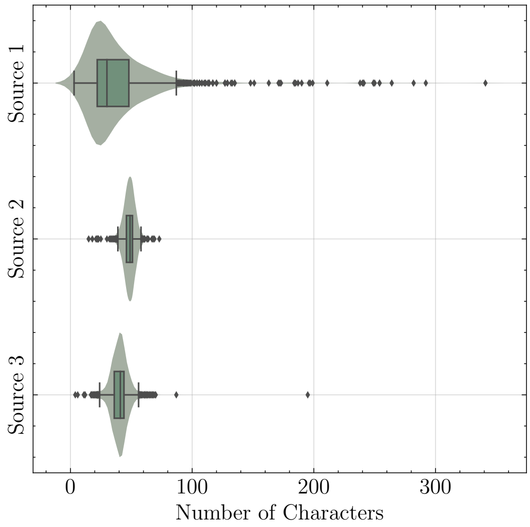


Figure 3: The distribution of label length by character count across the different dataset sources of HICMA.

the same style distributions, we relied on stratified sampling to preserve class distribution between the original set and produced subsets.

3.3 Dataset Preparation & Statistics

The data preparation process involved manually dividing the images into smaller segments. Images that originally contained multiple lines of text were further divided to create multiple images containing a single line of text. Images that only contained decorative motifs were discarded. This resulted in a total of 1,597 images from source 1, 1,480 images from source 2, and 1,954 images from source 3.

The combined HICMA dataset is thus made of exactly 5,031 images and is distributed across five styles: Kufic, Thuluth, Naskh, Diwani and Muhaquaq, with the Naskh style being the most prevalent followed by Thuluth as depicted in Figure 1. Figure 2 highlights that the most diverse set of calligraphy styles is present in source 1, whereas sources 2 and 3 predominantly consist of Naskh scriptures. This discrepancy in style diversity likely stems from the datasets' origins.

Source 1 encompasses a diverse collection of artistic Arabic calligraphy images, contributing to the wider variety of styles observed. In contrast, sources 2 and 3 comprise manuscripts only, where the Naskh style is mostly used for writing such scripts. The variation in style diversity is also evident in the sentence lengths within each set, as depicted in the violin plot in Figure 3. Although all three sets exhibit similar distributions of sentences with lengths under 100 characters and averaging around 50 characters, source 1 stands out due to the presence of numerous outliers with sentence lengths surpassing 300 characters.

The disparity in sentence lengths within source 1 can be explained by the nature of the images in this source. Calligraphy images allow for more text to be densely packed into a limited space compared to manuscript images. This aspect, combined with the challenge of segmenting intricate calligraphy words, contributes to difficulties in processing such images into smaller segments. For a visual representation refer to Table 2, which provides examples of images from all three dataset sources. The HICMA dataset is publicly available[7] for research purposes under the Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license.

4 Benchmark Experiments

4.1 Models

We investigated three state-of-the-art OCR tools that supported Arabic text recognition, namely TesseractOCR[8], Kraken (Kiessling, 2022), and EasyOCR[9], and describe them below. We ran the tools on the validation subset of the HICMA dataset (10%) for the presented benchmark evaluation.

1. **TesseractOCR**[8]: A widely-used open-source OCR engine developed by Hewelet-Packard and then by Google. It is a reliable and robust option for general text recognition tasks. The TesseractOCR engine is pre-trained for segmenting and recognizing text in images. Throughout our research, we assessed two pre-trained models for Arabic OCR from TesseractOCR[10] and ClearCypher[11].

2. **Kraken (Kiessling, 2022)**: An open-source tool specialized in recognizing historical and non-latin scripts, making it particularly suitable for the HICMA dataset. Kraken is trained on specialized datasets focusing on unique writing styles and scripts, allowing it to excel in scenarios where standard OCR engines might struggle. We evaluated the performance of three Kraken models pre-trained on Arabic manuscripts and publicly available online.

[7] https://hicma.net

[8] https://tesseract-ocr.github.io/

[9] https://www.jaided.ai/easyocr/documentation/

[10] https://github.com/tesseract-ocr/tessdata_best/blob/main/ara.traineddata

[11] https://github.com/ClearCypher/enhancing-tesseract-arabic-text-recognition

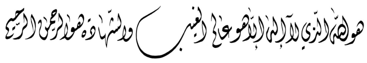
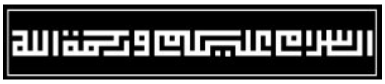
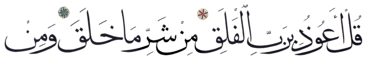
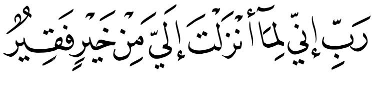
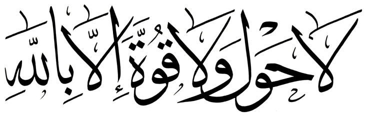
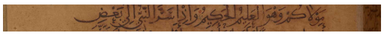
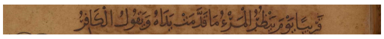
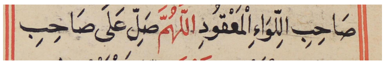
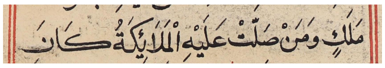
Image	Text label	Style	Source
	هو الله الذي لا اله الا هو عالم الغيب والشهادة هو الرحمن الرحيم	Diwani	Source 1
	السلام عليكم و رحمة الله	Kufic	Source 1
	قل أعوذ برب الفلق من شر ما خلق ومن	Muhaquaq	Source 1
	رب إني لما أنزلت إلي من خير فقير	Naskh	Source 1
	لا حول ولا قوة الا بالله	Thuluth	Source 1
	مولاكم وهو العليم الحكيم وإذ أسر النبي إلى بعض	Naskh	Source 2
	قريبا يوم ينظر المرء ما قدمت يداه ويقول الكافر	Naskh	Source 2
	صاحب اللواء المعقود اللهم صل على صاحب	Naskh	Source 3
	ملك و من صلت عليه الملائكة كان	Naskh	Source 3

Table 2: Sample images from HICMA along with associated labels, styles, and corresponding sources.

The three models will be referred to as Kraken-Arabic Best[12], Kraken-All Arabic Scripts[13], and Kraken-Arabic Generalized[14].

3. **EasyOCR**[9]: A user-friendly OCR library designed by Jaided AI that employs deep learning models to accurately segment and recognize text from images. It is designed to be easy to integrate into applications and supports multiple languages, including Arabic.

With the TesseractOCR and the Kraken models, the images were first transformed to grayscale and converted into binary format. In contrast, the images used for EasyOCR were not subjected to any pre-processing as no significant change in performance was observed. Moreover, as there were no available pre-trained Kraken segmentation models for Arabic, the images were resized to a smaller dimension of 200x1200 before being fed to the Kraken models. The image resizing helped decrease the inference time while also enhancing the accuracy of the Kraken models.

4.2 Evaluation Metrics

We utilized three evaluation metrics to assess the performance of the benchmark OCR models on the HICMA dataset.

1. **Levenshtein Ratio**: The Levenshtein Ratio (Sarkar et al., 2016) measures the similarity between two strings, that is, the ground

[12] https://zenodo.org/record/7050270/files/all_arabic_scripts.mlmodel

[13] https://zenodo.org/record/7050296/files/arabic_best.mlmodel

[14] https://github.com/OpenITI/OCR_GS_Data/blob/master/ara/abhath/arabic_generalized.mlmodel

	WER	CER	Levenshtein ratio
EasyOCR	**94.51%**	**58.47%**	**53.86%**
Kraken-Arabic Best	95.96%	65.84%	43.36%
Kraken-All Arabic Scripts	97.01%	67.14%	42.23%
Kraken-Arabic Generalized	100.55%	75.09%	34.82%
TesseractOCR-ClearCypher	98.99%	75.44%	31.94%
TesseractOCR	99.44%	81.96%	26.79%

Table 3: Summary of HICMA evaluation results across the three benchmark OCR models.

truth and OCR-generated text. It is derived from Levenshtein distance (Levenshtein, 1966), which calculates the minimum number of single-character edits required to convert one string into another and then computes the ratio of correct characters to the total number of characters in the ground truth text. A higher Levenshtein ratio reflects a more accurate OCR model.

2. **Character Error Rate (CER)** (Morris et al., 2004): The CER relies on the Levenshtein distance (Levenshtein, 1966) to calculate the ratio of incorrect characters recognized as compared to the ground truth text. It quantifies the accuracy of OCR models at the individual character level. The CER is associated with the portion of characters being incorrectly predicted. A lower CER reflects a more accurate OCR model with 0 being a perfect score. The CER score may exceed 1 if the value of insertions is high.

3. **Word Error Rate (WER)** (Morris et al., 2004): The WER calculates the ratio of incorrectly recognized words to the total ground truth words. Similarly to the CER, lower values of WER indicate better performance with 0 meaning the handwritten text was perfectly recognized. The WER may also exceed the value of 1.

All three metrics were developed using the python-levenshtein[15] package and are included in the benchmarking tool available on Github[16].

4.3 Model Results

Table 3 provides an overview of the models' performance on the HICMA validation set, measured using the three evaluation metrics: WER, CER, and Levenshtein ratio. Evidently, among the pre-trained models, the EasyOCR pre-trained model for Arabic text stands out in terms of performance. However, even the best-performing model falls short of meeting the requirements for a practical OCR system for handwritten text, as the standard acceptable character error rate is around 20%(Tomoiaga et al., 2019), a benchmark that these models are quite far from achieving.

A deeper examination of the EasyOCR model's performance, shown in Figure 4, reveals that it excels particularly in recognizing text written in the Naskh style. This style exhibits a CER that is 53% lower than Diwani, the next style in terms of performance. Furthermore, the Naskh WER is 7% lower while the Levenshtein ratio is 2 times higher than Diwani. The gradual decline in performance as we transition from Naskh to Diwani, Thuluth, Muhaqaq, and finally Kufic can be attributed to their frequency of usage as calligraphy fonts as present in our dataset as well as the characteristics of each style, making some more difficult to recognize than others.

Given that Naskh is one of the most commonly used styles for Arabic manuscripts and everyday writing, the success of the EasyOCR model in this style is expected due to its primary training on Arabic computer-generated text, using the Amiri and Noto Sans Arabic fonts[17]. These fonts are very similar to manuscript handwriting styles like Naskh. On the other hand, the remaining styles like Diwani, Thuluth, Muhaqaq, and Kufic are more ornamental and artistic in nature. Therefore, the model's accuracy diminishes in recognizing these artistic styles.

This variation in performance across different calligraphic styles highlights the significance of

[15] https://github.com/maxbachmann/Levenshtein

[16] https://github.com/anisdismail/HICMA-benchmark

[17] https://github.com/Belval/TextRecognitionDataGenerator

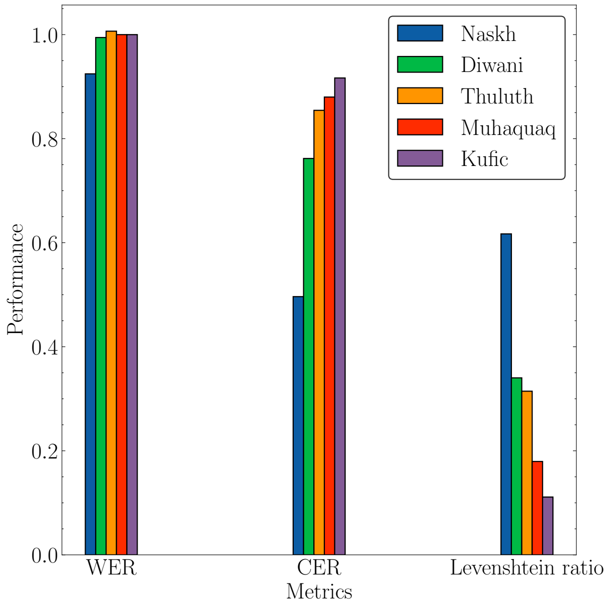


Figure 4: Performance metrics of the EasyOCR model across the different styles in HICMA.

having a diverse dataset that encompasses various styles. It also emphasizes the need to enhance OCR models' adaptability to challenging stylistic patterns within Arabic calligraphy. This endeavor would contribute to the development of more robust OCR systems capable of accurately recognizing text in images containing intricate calligraphy.

5 Limitations

As we present the HICMA Arabic dataset and the methodologies employed in this research, it is essential to acknowledge a few limitations that remain open for enhancement in future work.

- **Dataset Size and Style Diversity**: Despite HICMA being the most diverse public Arabic manuscript and calligraphy recognition dataset to date, there remains a need for further style diversification and an increase in sample count per text style. HICMA is currently composed from three sources, which do not represent the wide range of variations in Arabic texts. More so, the dataset's size remains limited compared to the vast range of Arabic texts available and would benefit from further expansion.

- **Pre-processing Challenges**: Given the inherent complexity of Arabic scripts and the variability in textual layouts, certain images in the HICMA dataset may present challenges during pre-processing. Some documents might contain lengthy texts or intricate structures, requiring manual segmentation or cropping and making it challenging to ensure reliable pre-processing across the dataset.

- **Model Limitations**: Variability in image quality, skewed perspectives, rotated motifs, and uncommon fonts have been shown to affect the existing OCR models' accuracy. To address existing Arabic OCR performance limitations, it is crucial to investigate the development of models that are fine-tuned to be native to Arabic manuscripts and calligraphy.

By addressing these limitations, future research will lead to advancements in Arabic OCR technology.

6 Conclusion

In this work, we presented HICMA as the largest and most diverse public dataset to date for Handwriting Identification of Calligraphy and Manuscripts in Arabic. The introduced dataset includes more than 5,000 images across five diverse Arabic text styles along with image-text sentence pairs and style labels for all images. This dataset fills the existing literature gap for Arabic manuscript and calligraphy text recognition. In this work, we detailed the data collection, labeling, and pre-processing steps of the created HICMA dataset. We further presented statistics about the dataset styles and label size diversity. We finally conducted a benchmark evaluation of the top three current state-of-the-art OCR models for Arabic and reported their performance on the HICMA dataset, serving as a baseline for future works. Upon analysis of the benchmark results, we highlight remaining open challenges in the HICMA dataset and the existing OCR models that support Arabic as a language. The HICMA dataset and the accompanied benchmarking tool are made publicly available for the research community. We believe our work is the first among many making more inclusive Arabic handwriting recognition for manuscripts and calligraphy possible.

7 Acknowledgements

We express our heartfelt gratitude to the dedicated volunteers and professors Vahid Behmardi and Lina Karam for their invaluable contributions to the dataset project. Special thanks for Hawraa Hellani, Kevin Nammour, Lyn Ismail, Nour Ismail, Pia

Kattoura, and all other volunteers for their meticulousness and attention to detail which significantly enhanced the dataset's quality. Their collective efforts exemplify collaboration, curiosity, and innovation, and without them, this project would not have been possible.

References

S Abdelazeem and E El-Sherif. 2017. The arabic handwritten digits databases: Adbase & madbase.

Kalthoum Adam, Somaya Al-Maadeed, and Ahmed Bouridane. 2017. based classification of arabic scripts style in ancient arabic manuscripts: Preliminary results. In *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, pages 95–98. IEEE.

Kalthoum Adam, Asim Baig, Somaya Al-Maadeed, Ahmed Bouridane, and Sherine El-Menshawy. 2018. Kertas: dataset for automatic dating of ancient arabic manuscripts. *International Journal on Document Analysis and Recognition (IJDAR)*, 21(4):283–290.

SR Allaf and R Al-Hmouz. 2016. Automatic recognition of artistic arabic calligraphy types. *Journal of King Abdulaziz University*, 27(1):3–17.

Bodour Alrehali, Najla Alsaedi, Hanan Alahmadi, and Nahla Abid. 2020. Historical arabic manuscripts text recognition using convolutional neural network. In *2020 6th Conference on Data Science and Machine Learning Applications (CDMA)*, pages 37–42. IEEE.

Najwa Altwaijry and Isra Al-Turaiki. 2021. Arabic handwriting recognition system using convolutional neural network. *Neural Computing and Applications*, 33(7):2249–2261.

Zaid Alyafeai, Maged S Al-shaibani, Mustafa Ghaleb, and Yousif Ahmed Al-Wajih. 2021. Calliar: An online handwritten dataset for arabic calligraphy. *arXiv preprint arXiv:2106.10745*.

Zineb Kaoudja, Mohammed Lamine Kherfi, and Belal Khaldi. 2019. An efficient multiple-classifier system for arabic calligraphy style recognition. In *2019 International Conference on Networking and Advanced Systems (ICNAS)*, pages 1–5. IEEE.

Manal M Khayyat and Lamiaa A Elrefaei. 2020. A deep learning based prediction of arabic manuscripts handwriting style. *Int. Arab J. Inf. Technol.*, 17(5):702–712.

Benjamin Kiessling. 2022. The Kraken OCR system.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707. ADS Bibcode: 1966SPhD...10..707L.

Sabri A. Mahmoud, Hamzah Luqman, Baligh M. Al-Helali, Galal BinMakhashen, and Mohammad Tanvir Parvez. Online-khatt: An open-vocabulary database for arabic online-text processing.

Sabri A Mahmoud, Hamzah Luqman, Baligh M Al-Helali, Galal BinMakhashen, and Mohammad Tanvir Parvez. 2018. Online-khatt: an open-vocabulary database for arabic online-text processing. *The Open Cybernetics & Systemics Journal*, 12(1).

Volker Märgner and Haikal El Abed. 2009. Icdar 2009 arabic handwriting recognition competition. In *2009 10th International Conference on Document Analysis and Recognition*, pages 1383–1387. IEEE.

Andrew Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition.

Sourour Njah, Badreddine Ben Nouma, Hala Bezine, and Adel M Alimi. 2012. Mayastroun: A multilanguage handwriting database. In *2012 International Conference on Frontiers in Handwriting Recognition*, pages 308–312. IEEE.

Seetah AL Salamah and Ross King. 2018. Towards the machine reading of arabic calligraphy: a letters dataset and corresponding corpus of text. In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, pages 19–23. IEEE.

Sandip Sarkar, Dipankar Das, Partha Pakray, and Alexander Gelbukh. 2016. JUNITMZ at SemEval-2016 Task 1: Identifying Semantic Similarity Using Levenshtein Ratio. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 702–705, San Diego, California. Association for Computational Linguistics.

Ciprian Tomoiaga, Paul Feng, Mathieu Salzmann, and Patrick Jayet. 2019. Field typing for improved recognition on heterogeneous handwritten forms. ArXiv:1909.10120 [cs].